nature

www.nature.com/nature

Vol 462 | Issue no. 7271 | 19 November 2009

The entangled bank unravels

This third special issue in *Nature*'s year-long celebration of Charles Darwin focuses on the dire challenges to Earth's biodiversity — and finds some reason for hope.

"It is interesting to contemplate an entangled bank, clothed with many plants of many kinds, with birds singing on the bushes, with various insects flitting about, and with worms crawling through the damp earth." So Charles Darwin begins the concluding paragraph of *On the Origin of Species*, published 150 years ago next week. By invoking this gentle image, Darwin sought to emphasize how "endless forms most beautiful and most wonderful" have all evolved through the process of natural selection.

Were he alive today, Darwin would have cause to be less rhapsodic. The modern version of his bank might well be dominated by invasive shrubs, having been denuded of most native plants by deforestation, and nearby streams would probably be polluted and filled with sediment from excess run-off.

It is hardly news that the rich pageant of life, which inspired Darwin and his work, is now suffering. According to data released this month by the International Union for Conservation of Nature in its Red List of Threatened Species, one-fifth of mammals and nearly one-third of amphibians are threatened with extinction, and the situation is no better among plants: almost one-third of known gymnosperms, the group that includes conifers, are threatened. Yet despite all the warnings from scientists and environmentalists, nations have done little more than fret over the problem. Although almost 200 countries have pledged through the Convention on Biological Diversity to significantly reduce the rate of biodiversity loss by next year, leaders of that effort acknowledge not only that the world will come up short of this target, but also that it was basically unachievable from the start and that it represented more of a political statement (see page 263).

This week, *Nature* ends its year-long celebration of Darwin (www.nature.com/darwin) by examining some of the most pressing issues concerning the loss of biodiversity, as well as ways to address the problem. The fact that upper levels of government are beginning to focus their attention on the biodiversity crisis gives some cause for optimism. For example, the United Nations General Assembly has named 2010 as the International Year of Biodiversity, with a meeting scheduled in New York next September at which heads of state will take up the issue. The following month, parties to the biodiversity convention will gather in Nagoya, Japan, to develop specific and verifiable biodiversity targets for nations over the coming decades. These meetings give countries an incentive to start protecting vital ecosystems during the next 11 months so that they can head to the Nagoya summit boasting of success.

There is growing recognition that diverse ecosystems can provide substantial economic benefits — a concept known as ecosystem services — which has strengthened support for conservation in the business and political communities. The News Feature on page 270 profiles ecologist Gretchen Daily of Stanford University in Palo Alto, California, an advocate of this concept who helped it to emerge as a major idea in conservation. Another article (page 266) shows this concept in action in Brazil, where it has helped to preserve the remaining patches of the species-rich Atlantic forest. And in an Opinion piece (page 277), the leader of an international study, known as the Economics of Ecosystems and Biodiversity (TEEB) project, argues that governments must put taxes and benefits in place to protect nature's 'public goods'. Just last week, the TEEB project announced initial results suggesting that investments in conservation can reap economic benefits that far exceed the initial outlay.

The situation in Brazil is a good example. Preserving patches of forest has not only helped the golden lion tamarin to survive, but has also helped to provide clean water, flood control and other economic benefits to nearby communities. These 'win–win' situations are natural starting points for conservation efforts because they are easily sold to politicians and other stakeholders.

Climate change will place new stresses on already weakened ecosystems but it can also present political and economic

For podcast and more online extras see www.nature.com/darwin

opportunities. One example is a strategy known as reducing emissions from deforestation and forest degradation (REDD). According to estimates by the Intergovernmental Panel on Climate Change, the clearing of forests accounts for approximately one-fifth of greenhouse-gas emissions by humans. Thus, stopping deforestation could be a relatively cheap and effective way to reduce emissions and slow the rate of global warming. At the same time, argue Will Turner and his colleagues in an Opinion piece on page 278, efforts to preserve natural ecosystems can help to ameliorate some of the effects of climate change. The international climate treaty currently under negotiation is likely to include a REDD mechanism that would provide funds to tropical countries to save their forests, a move that would help to mitigate climate change and sustain biodiversity.

Although ecosystem degradation looks set to increase in the future as a result of climate change, the biggest threat to biodiversity today is the rapid disappearance of habitats. At present, only around 14% of land surface and less than 6% of territorial seas are protected worldwide. Yet such areas help to support nearly one-sixth of the world's population, according to the TEEB study. As nations look beyond the likely failure of the 2010 biodiversity target, they should commit to placing more areas under protection. It will be crucial to select valuable sites that harbour the species that are most threatened. The wealthiest sectors of society tend to be the most removed from nature, whereas the world's poorest people rely heavily on the fruits of diverse ecosystems. As a result, care must be taken to ensure that conservation initiatives do not come at the expense of people, particularly indigenous communities that can be indirectly harmed when land is suddenly set aside. ■

Access denied?

Information-sharing resources are essential to biologists and deserve international support.

Every weekday, thousands of researchers around the world access the Arabidopsis Information Resource (TAIR), which contains the most reliable and up-to-date genomic information available on the most widely used model organism in the plant kingdom. But now, to those users' horror, TAIR faces collapse: the US National Science Foundation (NSF) is phasing out funding after 10 years as the data resource's sole supporter (see page 258).

TAIR's plight is emblematic of a broader crisis facing many of the world's biological databases and repositories. Research funding agencies recognize that such infrastructures are crucial to the ongoing conduct of science, yet few are willing to finance them indefinitely. Such agencies tend to support these resources during the development phase, but then expect them to find sustainable funding elsewhere.

Unfortunately, that is not easy. Other funding agencies are no more likely to provide long-term support than the agency that launched the resource in the first place. Moreover, any government agency's long-term plans are vulnerable to short-term political expediency. Witness, for example, Japan, where the new government has slashed the budget of the RIKEN BioResource Centre by one-third (see page 258).

Private firms are equally poor bets. Advertising and sponsorship are unlikely to bring in enough money to pay the experts needed to maintain such resources. And the superficially plausible idea of charging subscription fees is effectively unworkable for facilities such as TAIR, because the producers and consumers of data are essentially the same community. Scientists provide data and resources for free, because sharing benefits everyone. However, they would be considerably less likely to deposit the fruits of their labour if this synergy was removed from the equation. Subscription-based databases and resources would then enter a downward spiral, becoming less and less complete and so less and less valuable.

The problem is acute even for modest resources. Two examples are the kidney database EuReGene and the mouse-embryo database EURExpress, both of which were launched with funds from the European Commission that have run out in recent months. The databases are currently being maintained on a hand-to-mouth basis, and the scientists who built them don't know where to turn for maintenance money. Yet the European Commission's investment will have been wasted if the databases disappear.

It is time for a whole new approach. Front-line biology cannot function without these resources, so solutions must be found at both national and international levels.

Governments must ensure that at least one of their national funding agencies has money specifically set aside for the long-term support of bioresource infrastructures. A good model to emulate would be the United Kingdom's Biotechnology and Biological Sciences Research Council, which allows databases and other such resources to apply for ring-fenced funding, saving them from having to compete with hypothesis-driven grants, which are the agencies' mainstay.

"The sharing of bioresources does not and should not stop at national borders."

But action is also needed on the international front. The sharing of bioresources does not and should not stop at national borders. For example, only about a quarter of TAIR users are based in the United States. China is the second biggest user at around 12%, followed by Japan at around 10%. This is not atypical. Yet it is difficult for a single national agency to justify maintaining a resource for the rest of the world. What is required is an international cost-sharing organization that could fund competitively selected infrastructures, large and small.

The European Commission has made a good start with projects such as ELIXIR (European Life Sciences Infrastructure for Biological Information), which is studying ways of steering national agencies towards the joint funding of bioresources. A global, ELIXIR-like initiative is urgently needed, run perhaps by an international, relatively unbureaucratic organization such as the Human Frontier Science Program.

But an international solution may be a long time coming. In the meantime, bioresource infrastructures might be wise to invest some time in public relations, giving paymasters a greater understanding of the consequences of their decisions. ■

RESEARCH HIGHLIGHTS

G. ROUSE

Weird worms

BMC Biol. **7,** 74 (2009)

Feeding off whale bones at the bottom of the ocean, the *Osedax* genus of marine worms was first described by scientists in 2004. In these creatures, harems of tiny males are wholly encased in the tubes that surround the females. Now Robert Vrijenhoek of the Monterey Bay Aquarium Research Institute in Moss Landing, California, and his colleagues say there are at least another 12 putative species (*Osedax* orange-collar, pictured) in addition to the five previously described.

The team examined DNA sequences and physical traits of *Osedax* from whale remains and conclude that these newly discovered species have been evolutionarily separate for millions of years.

GEOLOGY

Impact ironed out

Geology **37,** 1011–1014 (2009)

A huge meteorite or comet that smashed into North America 1.85 billion years ago was responsible for the abrupt end of certain iron deposits in the rocks around Lake Superior, say John Slack and William Cannon from the US Geological Survey in Reston, Virginia. They propose that the collision, dubbed the Sudbury impact, caused dramatic changes in the oxygen levels of the deep oceans.

The impact probably caused a giant tsunami and other mixing processes that brought small amounts of dissolved oxygen to the previously oxygen-free deep ocean. Oxygen would have lowered the solubility of iron from hydrothermal vents, hindering its journey to the continental margin, an area where ocean crust and continental crust meet. This stopped the deposition of banded iron formations in the rocks of this region, say the authors.

CANCER BIOLOGY

Gene highs and lows

PLoS Genet. **5,** e1000719 (2009)

A large-scale survey of gene loss or gain in cervical cancer has flagged more than 50 potential genetic drivers of the disease.

Heidi Lyng and her colleagues at the Norwegian Radium Hospital in Oslo screened tumours from 102 patients with cervical cancer to look for changes in gene-copy numbers and expression profiles. They found 57 candidate genes that were frequently gained or lost and which were linked to various well-known tumour-promoting processes, such as carbohydrate metabolism and avoiding cell suicide. They also discovered some novel genes tied to resistance to chemoradiotherapy.

MICROSCOPY

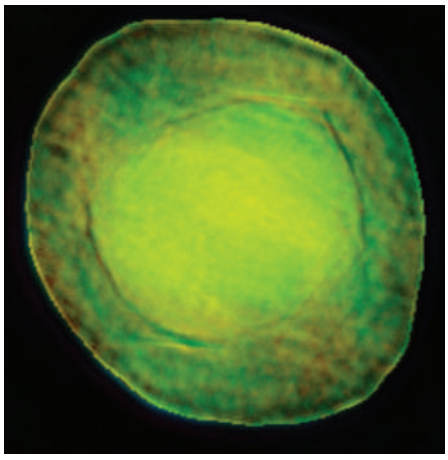
Cell close-up

Phys. Rev. Lett. **103,** 198101; 198102 (2009)

Researchers have taken the first X-ray diffraction images of intact, hydrated cells.

Because of their short wavelengths, X-rays can penetrate deep into specimens and generate high-resolution images, yet it has been difficult to use X-ray diffraction microscopy on intact cells because the radiation damages them. Freeze-drying the cells makes them more stable but they are still damaged after multiple exposures.

Chris Jacobsen at Stony Brook University in New York and his colleagues protected yeast cells (X-ray diffraction micrograph, pictured) from radiation damage by freezing them to below –170 °C. Because the cells were hydrated when frozen, their structures were similar to those of living cells.



AMERICAN PHYSICAL SOCIETY

Enju Lima and her colleagues at the European Synchrotron Radiation Facility in Grenoble, France, used a similar technique to image bacteria. Both groups were able to image the cells' internal structures at resolutions of less than 50 nanometres.

AGRICULTURE

Mixed manure message

Proc. R. Soc. B doi:10.1098/rspb.2009.16311 (2009)

Some organic-farming advocates have suggested that the nitrogen spike released from synthetic fertilizers attracts more insect pests to conventionally farmed crops. Joanna Staley at Imperial College London and her collaborators compared the abundance of insect pests on two sets of cabbage plots treated with either synthetic or organic fertilizer, including manure, over two seasons.

Different insects showed different preferences: one aphid species visited the organic cabbages more often in one year, but not in the other, while another aphid preferred synthetically fertilized plots, but only in one season. Such mixed results show that the impact of fertilization type on crop pests cannot be oversimplified.

PALAEONTOLOGY

Hot-blooded dinosaurs

PLoS ONE **4,** e7783 (2009)

Two methods for estimating animal metabolic rates have been applied to extinct dinosaurs to show which of the bipedal species may have been warm- or cold-blooded. Herman Pontzer of Washington

E. VAN BOMMEL

University in St Louis, Missouri, and his colleagues tested the models on 14 species, from small bird-like creatures to the large tyrannosaurus. The models, which predict metabolic rates on the basis of anatomy and the energetic cost of walking or running, suggest that the five largest species — each weighing more than 20 kilograms — had a higher metabolic output and were thus warm-blooded. Only the smallest species tested, a 0.25-kilogram *Archaeopteryx*, was deemed to be cold-blooded.

NANOBIOTECHNOLOGY

Magnetic tumour cells

Nature Nanotechnol. doi: 10.1038/nnano.2009.333 (2009)

Cancer can become more deadly when tumour cells spread from one tissue type to another. A technique that uses two kinds of nanoparticle could help to trap and detect these rare cells in the bloodstream, potentially enabling earlier cancer diagnosis, according to Vladimir Zharov of the University of Arkansas for Medical Sciences in Little Rock and his colleagues.

The researchers first inoculated the flank areas of mice with human breast-cancer cells to create tumours, followed later by an injected mixture of magnetic iron nanoparticles and gold-plated carbon nanotubes, each of which could bind to different cancer-cell receptors. The team then applied a magnet to the mouse ear to capture the particle-bound tumour cells, and pulsed the particles with a laser that triggered a photoacoustic signal. The technique detected a rising number of tumour cells in the blood vessels as the primary tumour developed throughout the four-week experiment.

MARINE ECOLOGY

Speedy sponge

J. Exp. Biol. **212,** 3892–3900 (2009)

The marine sponge *Halisarca caerulea* takes up about two-thirds of its body weight in dissolved carbon each day by filtering massive amounts of water, but it barely grows in size.

To work out where the carbon is going, Jasper De Goeij of the Royal Netherlands Institute for Sea Research in Texel and his colleagues collected the sponges from a coral reef (pictured, above) off the Caribbean island of Curaçao in the Netherlands Antilles and stained them with chemicals that reveal actively dividing and dying cells.

They found that some of the sponge's cells divide about every five hours — remarkably fast for a multicellular organism. Most of these fast-dividing cells were from the sponge's filtering chambers. The researchers did not observe much cell death but found that the creatures shed huge amounts of these cells, resulting in the conversion of dissolved carbon into a food source for other reef organisms.

For a longer story on this research, see go.nature.com/t2tgg7

PLANETARY SCIENCE

Cracking Martian ice

Geophys. Res. Lett. **36,** L21203 doi:10.1029/2009GL040634 (2009)

After nearly six months of scraping at the polar regions of Mars, NASA's Phoenix lander in 2008 discovered both ice-cemented soil and slabs of pure ice below the surface. It also found that thermal contractions had cracked the ground surface into polygons. Many on the Phoenix team had attributed the cracking to churning of the surface soil layer caused by seasonal frost cycles.

But by comparing geological features from cold, dry regions of Antarctica to those seen in Phoenix photographs, Joseph Levy of Portland State University in Oregon and his colleagues conclude that sublimation of the deeper slab ice is a better explanation for the cracking. The researchers surmise that the Martian landscape has been chiselled more by the steady loss of massive slab ice from below than by frost action in the surface soil.

REGENERATIVE BIOLOGY

Brainy stem cells

Proc. Natl Acad. Sci. USA **106,** 19150–19155 (2009)

Patients undergoing irradiation for brain tumours often display signs of cognitive dysfunction, owing in part to the loss of healthy neural stem and precursor cells. To investigate possible treatments, Charles Limoli of the University of California, Irvine, and his colleagues injected human embryonic stem cells into the brains of irradiated rats. After four months, the researchers confirmed the cells' survival in the rats' brains and found that the animals performed much better in a place-recognition task compared with irradiated rats that did not receive the transplant.

JOURNAL CLUB

Matt Friedman
University of Oxford, UK

A palaeontologist ponders how biodiversity is spread across the vertebrate tree of life.

Why do some biological groups burst at the seams with many different species, whereas others, despite their deep evolutionary heritage, contain only a handful of members? Many of my old vertebrate-biology textbooks are rife with qualitative scenarios, peddled with surprising degrees of confidence, that explain how species-rich branches can chalk up their success to key evolutionary 'innovations' and how less-diverse ones haven't kept up with changing conditions. What you won't find are details of how these exceptional groups might be identified in the first place.

Michael Alfaro of the University of California, Los Angeles, and his colleagues have now quantified this black art (*Proc. Natl Acad. Sci. USA* **106,** 13410–13414; 2009). They marry statistically explicit models with fossil-calibrated evolutionary trees and counts of living species to ask a basic, but surprisingly unanswered question: precisely which branches of the vertebrate family tree are more or less species-rich than expected given their age?

The authors identify nine groups that show substantial changes from the background tempo of vertebrate evolution: 'living fossils' such as lungfishes are characterized by lower-than-predicted diversity, whereas other branches, such as the perch-like fishes and a subset of mammals, contain vastly more species than expected.

As a palaeontologist, I am intrigued that three of the exceptionally diverse radiations are thought (although not without controversy) to have proliferated following the mass extinction that killed off the dinosaurs, hinting at the far-reaching consequences of this event in structuring the modern vertebrate fauna. Most importantly, these authors establish a clear quantitative framework that can be used to test all those textbook stories. I'm confident that in a few years, my students will learn a much more nuanced picture of vertebrate diversification than I ever did, one that will trace its own roots back to studies such as this.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

NEWS BRIEFING

POLICY

DNA bar codes: A combination of two gene regions, known as *rbcL* and *matK*, will be used as a 'bar code' to uniquely identify every species of land plant, biologists announced last week at the Third International Barcode of Life Conference in Mexico City. The two-gene identifier beat a pair of other proposals put forward by the 52-member plant working group of the Consortium for the Barcode of Life in July (see go.nature.com/nzTUhW). The panel plans to re-evaluate the decision in 18 months.

Climate stand: The American Physical Society (APS) last week rejected a call from some of its members to reverse its position on climate change. The society's 2007 statement acknowledges anthropogenic global warming. But a petition, delivered in July and now signed by more than 200 people, including Nobel laureate Ivar Giaever, had asked the APS to adopt a statement that climate change is a natural phenomenon. After a review, the society's council decided to retain its current statement. A separate committee will examine whether future "improvements in clarity and tone" can be made.

Italian reform: The Italian government has approved a draft law granting public research organizations greater independence. Under the changes, bodies such as the National Research Council, the National Institute of Nuclear Physics and the space agency, will be able to write their own statutes and regulations. If parliament approves the law, which it may do by the end of the year, all the institutes will get new heads, selected for the first time by committee, rather than by direct government appointment. Under the law, a small proportion of institutional research funds will be distributed according to merit, but no new money will accompany the reform.

PELICAN RECOVERY

The brown pelican (*Pelecanus occidentalis*, pictured) is being removed from a list of threatened and endangered species under the US Endangered Species Act after the government declared it officially recovered. Pelican populations were devastated by hunting, habitat destruction and the pesticide DDT, but the US Fish and Wildlife Service says there are now more than 650,000 in the United States, the Caribbean and Latin America. So far some 20 species have been delisted after recovering. Nine of these have been in the twenty-first century, including the bald eagle (*Haliaeetus leucocephalus*) and the Northern Rocky Mountain grey wolf (*Canis lupus*).

US FISH AND WILDLIFE SERVICE

Ecosystem economics: Countries will gain huge fiscal returns by protecting and restoring ecosystems, according to a 13 November report aimed at policy-makers and backed by the United Nations Environment Programme. The Economics of Ecosystems and Biodiversity (TEEB) study pointed to the ample financial returns of investment in protecting natural areas such as mangroves, tropical forests and grasslands. It also advocated cutting subsidies for environmentally harmful fossil fuels. See also page 277.

Security screen: Five gene-synthesis companies in a new International Gene Synthesis Consortium have adopted practices that are intended to address the biosecurity risks of the technology. The consortium's members will screen incoming orders against a single database — still being developed — that contains gene sequences "identified as potentially hazardous by authoritative groups", such as those within the European and US governments. The consortium will compete for members with the International Association of Synthetic Biology, based in Heidelberg, Germany, which released its own code of conduct earlier this month (see *Nature* doi:10.1038/news.2009.1065; 2009).

SOUND BITES

"I couldn't find a professional job in my chosen field because I didn't have my PhD yet."

Brooke Magnanti, now a cancer epidemiologist at the Bristol Initiative for Research of Child Health, UK, reveals that she was the anonymous sex worker and blogger Belle de Jour. Magnanti says she worked for an escort agency, charging £300 (US$500) an hour, after running into financial difficulties during her PhD.

Source: The Sunday Times

Censorship row: Australia's national science agency has sought to defuse accusations that it is gagging scientists by allowing the publication, after some rewording, of a paper critical of the effectiveness of cap-and-trade schemes in controlling carbon emissions. The paper, by Clive Spash, an ecological economist at the Commonwealth Scientific and Industrial Research Organization (CSIRO) in Canberra, was accepted by the journal *New Political Economy* earlier this year but then withdrawn from publication by the acting chief of his division. The agency had said that the article breached CSIRO rules by commenting on government policy — a charge Spash denies. See go.nature.com/bCusi6 for more.

FUNDING

Cash squeeze: The board of the Global Fund to Fight AIDS, Tuberculosis and Malaria, the major funding channel for

controlling these diseases, last week approved US$2.4 billion in extra funding over two years. The money includes $2 billion for the fund's main focus — proposals put forward by affected countries themselves — but is still $200 million short of that requested. The meeting, held in Addis Ababa on 12 November, was the agency's ninth funding round since its creation in 2002, and brings the total sum it has disbursed to $18.4 billion.

RESEARCH

Lunar splashdown: A NASA probe sent crashing into the Cabeus crater near the Moon's north pole on 9 October ploughed up a plume containing water, hydrocarbons and, unexpectedly, mercury, the agency said last week. Parts of the crater remain in permanent shadow, and so contain a record of the Solar System's chemistry and evolution because material that falls in freezes, becoming trapped. The water vapour and hydrocarbons detected by the Lunar Crater Observation and Sensing Satellite could have reached the Moon through impacts with comets rich in organic compounds. See go.nature.com/oDK7he for more.

Carbon cutters: Brazil has pledged to reduce its projected carbon dioxide emissions in 2020 by 36–39% below business-as-usual levels, increasing pressure on other countries less than a month before the United Nations climate summit in Copenhagen. The voluntary commitment builds on an existing pledge to cut the rate of deforestation by 80% by 2020; the government announced last week that roughly 7,000 square kilometres of forest were cleared this year, a drop of about 45% from last year's levels.

University unrest: Student protests against tuition fees, overcrowded courses and excessive workloads in newly established bachelor-degree programmes spread last week to 20 German cities. On 12 November, universities in Berlin and Tübingen called in police to evict students who were occupying lecture halls, although no violence was reported. German education minister Annette Schavan called on the country's federal states to streamline degree requirements. In Austria, where students have been demonstrating against overcrowded courses for weeks, pressure mounted to limit an influx of German students facing tuition fees and restricted admission at home.

Pathogen negligence: Canada's government laboratories are doing a poor job of keeping track of some pathogens, according to an audit by the Public Health Agency of Canada. The audit found that the labs' inconsistent tracking systems — a mixture of manual and electronic recording — might result in a pathogen being lost or used inappropriately. The audit included the National Microbiology Laboratory in Winnipeg, which handles samples of H1N1 pandemic flu, and which earlier this year lost track of 22 vials containing harmless Ebola-virus genetic material. Tracking systems for the most dangerous pathogens are more rigorous but could still be improved, the audit found.

NEWS MAKER

KIM JAE-HWAN/AP

Lee Myung-bak

The president of South Korea is backing a plan to increase total R&D spending to 5% of GDP by 2013. The country will also pump US$865 million into materials science. See go.nature.com/K3gYtx for more.

AWARDS

Youth-development prize: Laurence Steinberg, a psychologist at Temple University in Philadelphia, Pennsylvania, has been awarded the first Klaus J. Jacobs Research Prize for Productive Youth Development. Steinberg's interests include brain development in adolescents and its implications for juvenile criminal justice. He received the prize of 1 million Swiss francs (US$985,000) from the Jacobs Foundation, based in Zurich, Switzerland, which was set up by the late chocolate and coffee magnate Klaus Jacobs. The award, which will be presented on 3 December, must fund research.

THE WEEK AHEAD

25 NOVEMBER

In a combined event to be held in Washington DC and London, scientists and politicians will release findings on the public-health effects of policies to reduce greenhouse-gas emissions.

go.nature.com/4vFwwo

25–27 NOVEMBER

Planetary scientists will discuss observations and lay future plans for studying the methane detected on Mars, at a workshop at the European Space Agency's centre for Earth observation in Frascati, Italy.

www.congrex.nl/09c26

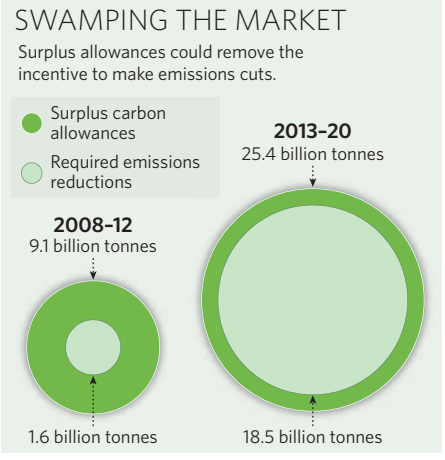
21 NOVEMBER

Part of the Genetic Information Nondiscrimination Act takes effect in the United States. The act, which was signed into law on 21 May 2008, prohibits discrimination in employment and health-insurance coverage on the basis of genetic information.

go.nature.com/VlYm5n

BUSINESS WATCH

One of the questions facing climate negotiators in Copenhagen next month is how to handle surplus carbon allowances in Russia and the former Eastern Bloc countries whose economies collapsed after the break-up of the Soviet Union. The Oslo-based consultancy Point Carbon projects that surplus carbon allowances will add up to the equivalent of 9 billion tonnes of carbon dioxide during the period 2008–12. That is nearly six times higher than emissions reductions required under the Kyoto Protocol. This excess of allowances looks set to continue. Point Carbon also analysed a 2013–20 scenario that takes into account the economic downturn and current commitments by developed countries. It projected an additional surplus of 16.3 billion tonnes, which increases to 25.4 billion tonnes if surplus allowances are carried over from the period to 2012. That compares with pledged emissions reductions of 18.5 billion tonnes. In theory, these surplus allowances could eliminate any incentive to reduce emissions. Few countries have been willing to buy credits, however, because they do not represent new greenhouse-gas reductions. European Union officials would like to eliminate them altogether in a new global-warming treaty, but countries holding the allowances could push for some kind of a compromise that maintains a discounted value going forwards.



NEWS

Japanese science faces deep cuts

The government's election promises vowed more support for science, but so far budgets look set to shrink.

Japanese researchers are in uproar about the drastic budget cuts being recommended for science projects by a new cabinet-level government advisory unit.

Since 11 November, working groups of the Government Revitalization Unit, created in September and chaired by Prime Minister Yukio Hatoyama, have been re-evaluating 220 government-funded programmes, including dozens of prominent science projects.

The drastic shake-up will hit the SPring-8 synchrotron in Harima, a planned supercomputer that was destined to be the world's fastest, ocean drilling projects and basic grant programmes, to name but a few.

The recommendations, part of an effort to trim ¥3 trillion (US$ 33.7 billion) off next year's budget, are the most concrete indication so far that Japan's new government intends to make comprehensive, long-lasting changes to the country's research priorities.

Scientists are reacting with frustration and, in some cases, apocalyptic predictions. One prominent crystallographer, who requested anonymity, told *Nature*: "If this goes on, Japanese scientists, including young scientists, will flow overseas, and Japanese science will die."

Hatoyama's government rode into power in August, promising to shift government expenditure from wasteful projects to initiatives that will benefit the average person, such as ending highway tolls. In August, Hatoyama told *Nature* that he would nonetheless increase support for science[1].

But since then, his government has been slicing into budgets. In October, the science and education ministry reduced the total grants for 30 of the projects under the Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST) from ¥270 billion to ¥100 billion[2].

On 8 October, after chairing a meeting of the Council for Science and Technology Policy, Japan's highest science-policy body, Hatoyama noted that his cabinet is "extremely rare" because it includes several engineers, such as himself. "Because we too did research, we know that researchers and academics can get drunk on their own studies," he said, according to the economic newspaper *Nihon Keizai Shimbun*. "Isn't it more appropriate to promote research that matches a new social system?"

At daily hearings in Tokyo, the unit's three working groups are devoting one hour to each project under review. The sessions can be viewed live on the Internet[3], and recommendations for the latest projects to be evaluated are uploaded to the website daily, with the basic message displayed in red. This is a startling amount of transparency for Japan, where budgets are usually delivered after bureaucrats strike deals in back rooms. "It's difficult to cut deals now," says Atsushi Sunami, director of science and technology policy at the National Graduate Institute for Policy Studies in Tokyo.

The 19 members of Working Group 3, which is reviewing science projects, include economists, a financial strategist, local government officials and other representatives of the public, along with a few scientists. It is usually ministry officials, not scientists, who have had to defend the projects under review.

The working group has already recommended that the ¥10.8 billion annual budget of SPring-8, the merits of which "were not adequately explained", be cut by one-third to one-half and be supplemented by charging users.

"The cuts to SPring-8 are devastating," says

Plant genetics database at risk as funds run dry

The world's most valued plant database faces extinction because its funding is being phased out by the US National Science Foundation (NSF), and no alternative source is on the horizon.

"This is the wrong way to go," says genomics researcher Ernest Retzel of the National Center for Genome Resources in Santa Fe, New Mexico. "I believe it will set the field back."

The NSF says that it does not have a policy to support long-term, established research-infrastructure projects such as the Arabidopsis Information Resource (TAIR), which maintains a free, open-access database of genetic and molecular-biology data for *Arabidopsis thaliana*, or thale cress, the widely used model plant. "We didn't approach this decision in isolation, we considered our whole portfolio," says Peter Arzberger, director of the Division of Biological Infrastructure at the NSF. "We rely on peer review in setting our priorities." The NSF has suggested that TAIR develop its own self-supporting funding model, based on user subscriptions and other sources of income.

J. BURGESS/SPL

The popular model plant *Arabidopsis thaliana*.

But TAIR director Eva Huala told an international meeting on database and bioresource sustainability, held in Rome on 11–12 November, that introducing a subscription system would destroy, not save, TAIR.

Huala, a member of the Department of Plant Biology at the Carnegie Institution for Science in Stanford, California, presented preliminary results of a survey among TAIR users, which revealed that many would be reluctant to submit data to TAIR if these were not freely shared.

Established ten years ago, TAIR integrates data submitted by the community with data extracted from the literature, and it has evolved into the plant community's foremost authority on matters relating to plant genomics, regulating nomenclature and developing curation standards. It is much more widely used than other plant databases because of its all-inclusive nature and the quality of its curation.

TAIR also feeds information into other specialist databases, such as those of the National Center for Biotechnology Information in Bethesda, Maryland, and the international protein database UniProt. In addition, it links to the Arabidopsis Biological Resource

SPRING-8/RIKEN/JASRI

The SPring-8 synchrotron facility faces proposed cuts of one-third to one-half of its budget.

structural biologist Soichi Wakatsuki, director of the KEK Photon Factory in Tsukuba and a collaborator with SPring-8. "There's no other synchrotron in the world that is supposed to earn so much of its own income." He laments the review process as "one-sided", adding that researchers are given "no real chance" to defend their projects. Tomitake Tsukihara, a crystallographer at the University of Hyogo, adds that protein crystallography and other basic science done at SPring-8 will suffer, and is organizing a protest in response to the recommendation.

A supercomputer planned by RIKEN, Japan's network of research labs, had already been thrown into confusion by the sudden departure of electronics giants NEC and Hitachi from the project earlier this year[4]. The project should now be "virtually eliminated", says the working group, which saw no need for Japan to host the world's fastest supercomputer.

Other recommendations made by the group include slashing the funding for RIKEN's BioResource Center and its Plant Science Center, with budget cuts of one-third proposed for each; cutting Japan's deep-sea-drilling programme by 10–20%; and at least halving the budget for the Institute for Research on Earth Evolution in Yokosuka. In addition, various competitive grant programmes, including the Grants-in-Aid programme — the bread and butter of most researchers — should be "simplified and reduced". Further recommendations on a prototype Japanese–European fusion reactor, planned as part of the international ITER project to prove atomic fusion as a power source, were expected as *Nature* went to press.

Asked whether the proposed cuts contradict earlier pledges to increase scientific funding, or whether increases in funding to other fields will offset these proposed cuts, a representative for Hatoyama said that these issues were "under discussion".

The working groups' recommendations will be considered by the Government Revitalization Unit before being submitted to the finance ministry, which will announce its budget in late December. ■

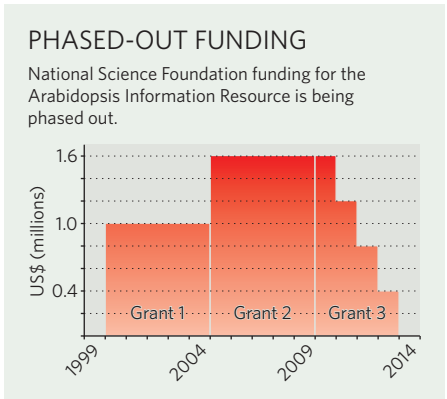
David Cyranoski

1. Cyranoski, D. *Nature* **460,** 938 (2009).
2. Cyranoski, D. *Nature* **461,** 854–855 (2009).
3. http://www.cao.go.jp/sasshin/
4. Cyranoski, D. *Nature* doi:10.1038/news.2009.495 (2009).

Center in Columbus, Ohio, which provides seed and DNA resources to researchers.

TAIR has been supported by two consecutive five-year NSF grants, the second of which came to an end on 31 August. The NSF is planning to maintain the current budget of $1.6 million for 2010, and then to phase out funding over the following three years (see graph).

TAIR



Huala's survey, sent on 4 November to more than 900 TAIR users, asked respondents which of TAIR's features are most important and which could feasibly be sacrificed in the hunt for alternative funding mechanisms. The majority of the 250 or so responses she has received so far say that there should be no log-in requirement, that everyone should have equal, free access to the data, and that data should continue to flow freely into other databases. Around two-thirds said they would be less likely to submit data to TAIR if these were not then freely shared with all researchers.

Respondents said that they would accept a situation in which publicly funded institutions and individuals had free access, but companies were required to buy subscriptions. They also said that they would be happy for advertising to appear on the website to raise revenue. But Huala says that neither measure would raise enough money to sustain the database.

"As soon as we introduce any form of subscription, we would not be able to export information to other free, open-access databases as we do now," she says. "The whole system would break down."

If TAIR were lost, another free database would inevitably spring up to take its place, thereby fragmenting the community, she adds. "The self-supporting models proposed by the NSF are more like a track to extinction."

Huala has put out feelers to other potential funding sources in the United States, and also to other countries, because only around a quarter of TAIR users are US-based. "But these are all long shots," she admits. "I am not optimistic."

Along with the Multinational Arabidopsis Steering Committee, which represents *Arabidopsis* researchers, the NSF is planning workshops for 2010 to gather input from members of the *Arabidopsis* community on their database and informatics needs.

The plight of TAIR is the most recent and most drastic example of funding crises now facing many databases and bioresources, says Paul Schofield, a molecular geneticist at the University of Cambridge, UK, who coordinated the Rome meeting. "There is a disparity between what science needs and available funding instruments for infrastructure," he says. "National research agencies need to get together to design new strategies." ■

Alison Abbott

See Editorial, page 252.

Muon collider gains momentum

Fermilab pins hopes on untested technology in race to stay at the cutting edge of physics.

As beams of protons again begin zipping around the underground ring of the Large Hadron Collider (LHC), teams of physicists are already competing to design a successor. Last week, US scientists staked their claim in a daring new venture: the world's first muon collider.

The collider could overtake two more-mature concepts, each of which plan to smash together electrons and positrons that have been accelerated through long, straight tunnels. But some physicists at the Fermi National Accelerator Laboratory (Fermilab) in Batavia, Illinois, are concerned about the expense and feasibility of the linear colliders, and question whether they would push the boundaries of physics beyond what the LHC is expected to achieve. They are now trying to rally enthusiasm for a collider that smashes muons, a particle that is about 200 times more massive than the electron.

"The question is, 'can we build it?'," said Fermilab director Pier Oddone at a muon-collider workshop held at Fermilab on 10–12 November. The meeting is seen as an important step in reviving an approach that has been in hibernation since it was first proposed in the mid-1990s. It also puts the muon collider in direct competition with the 31-kilometre-long International Linear Collider (ILC), a global design effort split between the United States, Europe and Asia, and the 48-kilometre-long Compact Linear Collider (CLIC) project based at CERN, the European particle-physics centre near Geneva, Switzerland, which already hosts the LHC.

Although it poses daunting research and development challenges, a muon collider would have many advantages over an electron smasher. The heavy particles offer a wider target and make for cleaner collisions, for example. Their ability to generate Higgs bosons, the as-yet unobserved particles believed to be part of the mechanism that gives particles mass, is expected to be so prodigious that muon colliders are sometimes dubbed 'Higgs factories'. But they have to accelerate the muons to near-light speeds quickly, so that the time-dilation effects of general relativity increase the lifetime of the particles beyond the typical 2.2 microseconds.

Muons can be boosted to these speeds in a relatively small — and therefore less expensive — accelerator ring because they don't lose much energy when muscled magnetically around curves. Circulating electrons emit much more energy, in the form of illuminating X-rays called synchrotron radiation, which is often used for crystallography.

Electron accelerators that need to reach energies of greater than a few hundred gigaelectronvolts (10^9 electronvolts) can limit their energy loss through synchrotron radiation by keeping the electrons moving in a straight line. But this makes for bigger, more expensive colliders (see graphic). A ring collider also has the advantage of higher collision rates, as the circulating particles get multiple chances to hit their target. Electrons in a linear collider get just one chance to hit the oncoming positrons, making precise beam alignment much more critical.

On the campaign trail

Just a few years ago, Oddone was lobbying to host the ILC (see *Nature* **435,** 728-729; 2005). Now, he favours the muon collider, partly because it would be small enough to fit comfortably on Fermilab's campus. The facility already hosts the world' highest-energy particle accelerator, the Tevatron — at least until the LHC reaches its target energy of 14 teraelectronvolts (10^{12} electronvolts). Moreover, Fermilab will soon have a way to make its own muons. After the Tevatron shuts down in a year or two, the lab intends to pursue Project X, an intense proton beam that will be used initially to study neutrinos, but could also be fashioned to make muons. A step-wise programme to a muon collider, says Oddone, makes more sense then biting off the "whole enchilada" at once.

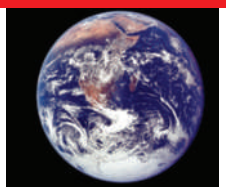
Fermilab has submitted a proposal to the US Department of Energy to form a national muon-accelerator project and to increase funding for muon-collider research from $9 million to $15 million per year. At a high-energy physics advisory meeting in Washington DC on 22 October, William Brinkman, director of the department's Office of Science, said that the high cost of the ILC has put it on the "back burner". Instead, he wants Fermilab to pursue the idea of a muon collider. "It would be nice to head in a direction that has some real innovation," he said.

Muon-collider advocates acknowledge that they have much catching up to do before they can compete with the ILC and CLIC, and Oddone wants to have a more detailed design concept in place by 2012. By that time, the LHC should have produced results revealing how much energy these colliders would need to reveal new physics.

The leader of the ILC effort, Barry Barish,

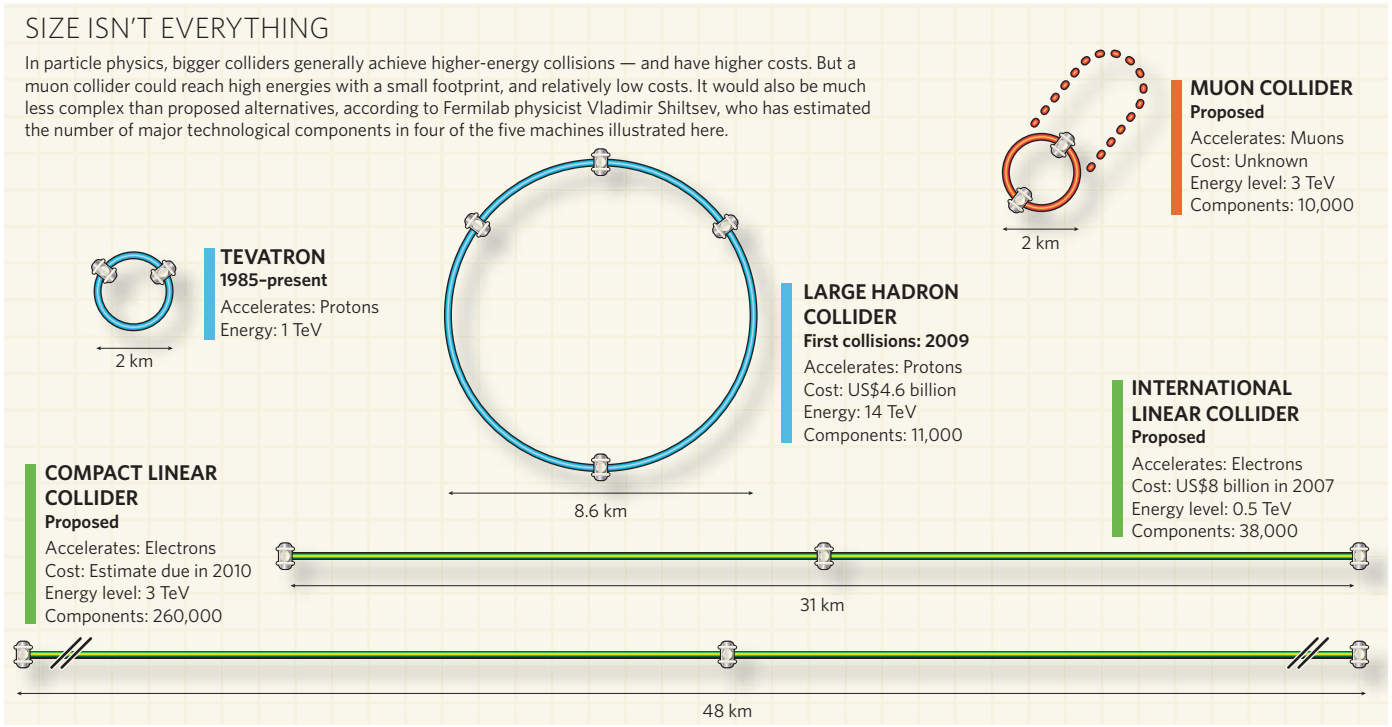
FERMILAB VISUAL MEDIA SERVICES

Fermilab (above) hopes to build a muon particle collider to keep it at the forefront of physics.



KEEPING EARTH COSY
Atmospheric nitrogen levels may have kept our planet habitable across eons.
go.nature.com/XjI9UI

NASA



says that muon-collider advocates now need to demonstrate that the technology can actually work. "The director of Fermilab has been campaigning for it," says Barish, who is also an emeritus professor at the California Institute of Technology in Pasadena. "That doesn't mean that anyone outside the United States takes it seriously."

Eye-watering costs

It's been a tough two years for the ILC, which was once the clear front runner among next-generation collider designs. It boasts a superconducting acceleration technology that has matured rapidly, and had an aggressive timetable set out. But at the end of 2007, budget cuts all but ended research efforts on the collider in the United States and Britain (see *Nature* **451,** 112–113; 2008).

The biggest problem with the ILC is its $20-billion price tag, says Robert Cahn, a physicist at Lawrence Berkeley National Laboratory in California who works at the LHC. He is blunt about the ILC's chances of being built. "The ILC is dead," he says.

Barish says that the project is still the most mature proposal, and that the true cost is closer to an $8 billion estimate made in 2007 than the energy department's projected figure of $20 billion, which reflects the price decades from now. He adds that the team is considering design changes that could cut the cost by 15%.

Barish acknowledges, however, that it would be a challenge to scale up the ILC beyond its target energy of 500 gigaelectronvolts if the LHC results suggest that its successor would need to explore even higher energy regimes. Building a bigger ILC to access those energies would require more superconducting cavities and more tunnel digging, a huge additional expense.

CLIC avoids the ILC's peak energy limitations with a nifty trick. Instead of relying on superconducting cavities, it uses conventional magnetic components, but in two beam tunnels. A beam in one tunnel would be accelerated until engineers slam on the brakes, siphoning off the energy and transferring it to an electron beam in the second tunnel. CLIC study leader Jean-Pierre Delahaye, based at CERN, declined to estimate costs for the project until a design study is finished next year. But it aims to achieve energies of 3 teraelectronvolts — six times the peak energy of the ILC — and has tunnels nearly 20 kilometres longer, making it a much more expensive prospect. CLIC will also require 440 megawatts of power, equivalent to the output of a small nuclear power station and much more than CERN has available.

"The director of Fermilab has been campaigning for it. That doesn't mean that anyone outside the United States takes a muon collider seriously."

The fate of the three concepts depends partly on the energy levels needed for the LHC to make its discoveries. Many theorists expect new high-energy particles to arise from supersymmetry, a theory that would double the number of known particles in the quantum bestiary. If the LHC doesn't find any, then the relatively low energies of the ILC might not be enough to uncover much of interest, further boosting the case for Fermilab to build a muon collider.

There are already indirect hints from experiments that rule out lower-energy supersymmetrical particles, says Fermilab's Steve Geer, who heads the muon-collider research and development effort. "It is all pointing to the energy scale for the new physics being higher."

However, even a 3-teraelectronvolt muon collider could reveal physics that is inaccessible to the LHC. The LHC smashes protons at much higher energies, but because those protons are made up of smaller particles, called quarks, collisions are much messier, cutting their effective energy by about a factor of ten compared with electron or muon colliders.

Despite their optimism, physicists such as Fermilab's Alan Bross, spokesperson for an experiment called MuCool (Muon Ionization Cooling Research and Development), must first solve the thorniest problem of a muon collider. MuCool is trying to work out how to get a hot cloud of muons to line up and march along in the cool, narrow beam needed for acceleration before all the muons decay.

Delahaye is glad that the muon research is being done, but he reminds his competitors that it can take a long time to win the case for a collider: he began working on CLIC in 1986. "To go immediately to the conclusion that [a muon collider] is simpler and less expensive, that's going too far," he says. ■

Eric Hand

K. CAMPBELL/GETTY

GOT A NEWS TIP?
Send any article ideas for *Nature*'s News section to **newstips@nature.com**

Efforts to sustain biodiversity fall short

But the issue is gaining attention as nations prepare for next year's summit.

With nations admitting that they will fail to achieve their goal of significantly cutting biodiversity loss by 2010, a flurry of work is under way to develop new, more robust targets and ways of monitoring progress. These must be ready by next October, when the 193 parties to the Convention on Biological Diversity (CBD) meet in Nagoya, Japan.

The summit will be the culmination of a series of events focused on the loss of biodiversity. The General Assembly of the United Nations has declared 2010 the International Year of Biodiversity, and governments will meet in London in January to start thrashing out a new set of biodiversity protections. In September, world leaders will convene in New York for a UN meeting on the topic.

In 2002, ten years after the CBD was signed, nations adopted a 2010 target in which they pledged to achieve "a significant reduction of the current rate of biodiversity loss at the global, regional and national level". But conservation leaders acknowledge that nations will come to Nagoya having fallen short. "We will not meet the target," says Matt Walpole, head of ecosystem assessment at the United Nations Environment Programme's World Conservation Monitoring Centre in Cambridge, UK. "Biodiversity is still in decline. This is because a lot of the key threats, such as change in land cover, habitat loss and pollution, are still not under control. And then new emerging threats, such as climate change, are coming through."

The parties to the convention are expected to set targets that are more specific than the current one. Nations could set a long-term goal for biodiversity in 2050, as well as an array of 2020 targets.

For example, Walpole suggests that targets could include ensuring that no new invasions by non-indigenous species take place. Another potential option would be setting a benchmark that 80–90% of all fisheries have to be be sustainable, he adds. The new targets will go beyond the broad efforts made in the 2010 goal to increase the amount of land and marine areas that are under protection. Goals could also include the requirement for governments to have plans for dealing with specific causes of biodiversity loss, such as deforestation.

A. MORRISON/TOLEDO BLADE/NEWSCOM

Construction of a dam in Tanzania helped drive the Kihansi spray toad to extinction in the wild.

"One of the challenges and weakness of the current target is that it is not sufficiently clearly defined," says Mike Parr, secretary of the Alliance for Zero Extinction, a partnership of conservation organizations. Parr says the alliance wants a goal of stopping extinctions to be included in the 2020 targets.

The biodiversity convention will also specify the metrics that parties must use to measure progress towards agreed goals. Examples include the Red List of Threatened Species, published by the International Union for Conservation of Nature (see 'Species at risk'), and the Living Planet Index, produced by the conservation group WWF and the Zoological Society of London's Institute of Zoology.

"The 2010 target was adopted without identifying the means of achieving the target. It was more a political statement," says Ahmed Djoghlaf, executive secretary of the CBD. He says the same mistake will not be made in Nagoya.

Ben Collen, a researcher at the Institute of Zoology, says that a lack of investment from governments in developing new indicators, such as measuring genetic diversity, has meant that important aspects of biodiversity are deficient in data. "The indicators we have at the moment will slot into the new framework, but new ones will also have to be developed," he says.

The CBD will report its final assessment of countries' progress towards the 2010 goal in May next year, when it publishes the third edition of its Global Biodiversity Outlook. The assessment will be drawn from reports submitted by countries that are party to the convention. Djoghlaf says that of the 86 national reports received so far, "all say, without exception, that they have failed to reach the target".

But behind the gloomy headline figures lie considerable efforts to improve the state of the world's biodiversity.

Since the 2010 target was adopted in 2002, the Brazilian government has increased the proportion of land designated as protected by 25% and deforestation rates have been reduced by 60%. It plans to identify further priority areas for conservation over the coming year.

And in Sweden, 9 new marine nature reserves were established between 2007 and 2008, bringing the nation's total to 21 sites. A further seven marine protected areas and six no-fishing areas are planned by 2010.

A surge of further efforts can be expected before next October to ensure that countries will be able to report some positive news. ■

Natasha Gilbert

See Editorial, page 251, and whole biodiversity special at www.nature.com/darwin.

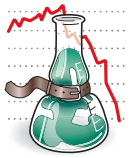
SPECIES AT RISK

Groups evaluated most completely	Number of described species	Percentage threatened
Gymnosperms	1,021	32%
Amphibians	6,433	29%
Mammals	5,490	21%
Birds	9,998	12%

Source: International Union for Conservation of Nature

Growth in R&D investment holding up

But analysts expect effects of financial crisis to be more marked next year.


RECESSION WATCH

Globally, corporate investment in research and development (R&D) grew more slowly in 2008 than in 2007, but remained relatively robust despite the economic crisis, according to a new report by the European Commission.

The 2009 European Union (EU) Industrial R&D Investment Scoreboard, released on 16 November, found that global corporate R&D investment rose by 6.9% in 2008, compared with 9.0% in 2007. EU companies saw their R&D investment grow by 8.1%, outpacing the 5.7% growth rate for US companies.

However, analysts say that the main impact of the financial crisis will probably be reflected more fully in next year's results.

"There are some first signs that firms have declining R&D expenditure because of cost; it's a weak decline, but I would expect a strong decline in 2009," says Hugo Hollanders, a senior researcher at the United Nations University-MERIT training centre in Maastricht, the Netherlands.

Meanwhile, the continued strong performance of European companies masked some more disquieting indicators for the EU. For several years, the region has led in R&D spending in 'medium-intensity' sectors, such as automobiles and parts, electronic and electrical equipment and chemicals, with the United States leading in 'high-intensity' R&D sectors, such as pharmaceuticals, biotechnology and information technology. That remains the case — but it is more marked than in the past.

"I was intrigued that the gap between high-tech and medium-tech in Europe and the United States has increased," says Ben Martin, professor of science and technology policy studies at the University of Sussex, UK. Contributing factors, he says, might include the increased pressures of globalization and competition, which often force countries to focus on the industries in which they already have an advantage.

Growing specialization could ultimately reverse the EU's overall lead in investment spending growth, says Hollanders, adding that "there are some concerns that there are not enough firms in R&D-intensive sectors in Europe". Switzerland, one of the most high-intensity R&D countries outside the EU and the United States, continues in its leading position, and its investment is growing faster than that of the EU — in part because of its main companies' specialization in the pharmaceutical and biotechnology sectors, he says.

"There are concerns that there are not enough firms in R&D-intensive sectors in Europe."

Although the report highlights the fact that growth in R&D investment remained at a reasonable pace in Europe and the United States in 2008, it notes that R&D intensity — defined as the investment by a company in R&D divided by its net sales — has effectively flatlined in firms in both regions over the past couple of years (see graph). Europe's lack of progress in this area makes it unlikely that it will fulfil the strategy outlined at the 2000 Lisbon summit, which committed member states to an R&D target of 3% of national gross domestic product by 2010.

"It's clear that the Lisbon target was never more than fanciful," Martin says. "It could only have been achieved if industry as well as government rapidly increased the amount of R&D spending."

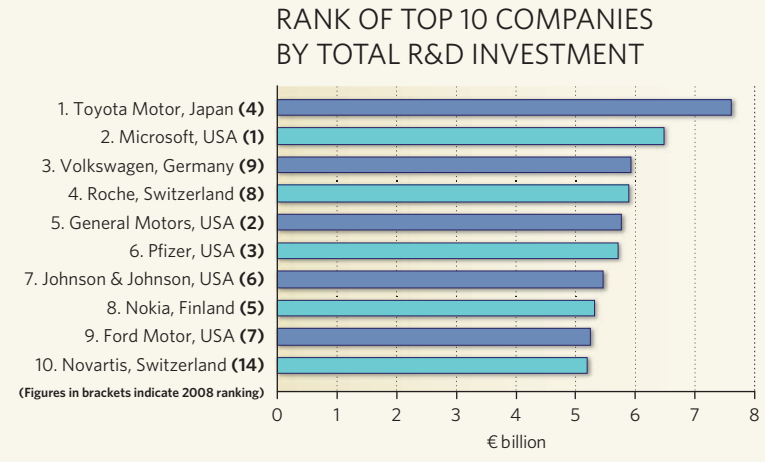
In another development that could ultimately shift regional balances, companies from emerging economies registered some of the highest R&D-investment growth rates despite accounting for a small share of total R&D. For instance, China posted an R&D growth rate of 40%, and India 27.3%. "These countries are rapidly becoming top players in R&D," says Martin.

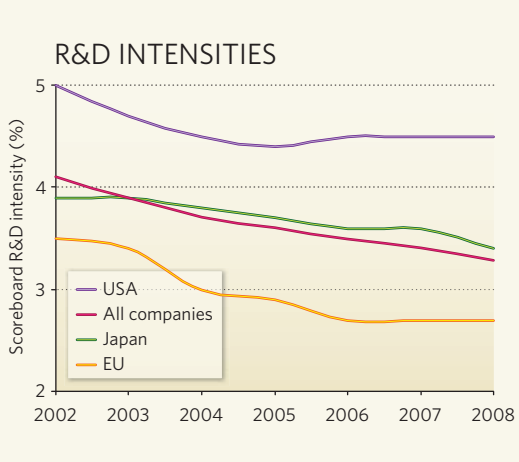
Hollanders, however, is more sceptical. "They are growing fast," he says, "but if you look at the levels they are coming from, even if the growth records are accurate, it will take a long time before they come up to the level of the EU or the United States."

Although the pharmaceutical, biotechnology and information-related sectors continued to dominate investment growth in both the EU and the United States, the automotive sector remained surprisingly strong, with car makers forming four of the top ten global R&D companies by total investment (see chart).

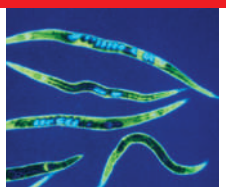
The scoreboard analysis includes only companies with publicly available audited accounts, and that itemize R&D expenditure in their financial reports, and it assigns companies to the country of their registered office, even if R&D operations are located in a different region. ■

Andrea Chipman





SOURCE: EUROPEAN COMMISSION



J. KING-HOLMES/SPL

BIOLOGISTS TURN AGAINST WORM
Researchers seek out alternative model organisms to *C. elegans*.
go.nature.com/PvZJNd

Fresh hope for German stem-cell patent case

The German federal supreme court has referred a controversial and already lengthy patent dispute about human embryonic stem (hES) cells to the European Court of Justice.

The move might drag out the case — first brought to the German patent court in 2004 — for another two years. But it could be worth the wait, says Oliver Brüstle, director of the Institute of Reconstructive Neurobiology at the University of Bonn and owner of the patent in question. He hopes that a ruling from the European court will finally settle some of the uncertainties that hamper stem-cell scientists in his country.

"It's crazy that patenting your methods is deemed to be contrary to public order."

German laws governing hES cell research are among the most restrictive in Europe. European patenting rules are meant to guide national legislation, but they include some statements that opponents and proponents of hES research can interpret in different ways — in particular, that patenting should not be allowed for procedures that are immoral or threaten public order.

The disputed patent was awarded to Brüstle by the German patent office in 1999. It covers a technique for generating nerve cells from established hES cell lines, which researchers in Germany are allowed to use. The method is a first step in generating neurons that could be used clinically to repair damage to the brain and spinal cord.

In 2004, Greenpeace lodged an objection to the patent on the grounds that the hES cell lines originated from fertilized human eggs and, as such, the patent offended public morality, threatened public order and contravened legislation that prohibits the industrial use of human embryos. In 2006 the federal patent court ruled in Greenpeace's favour. Brüstle then appealed the case by taking it to the federal supreme court.

Supreme-court judge Peter Meier-Beck said on 12 November that he needed more legal clarity before his court could make a decision. Because German patenting laws for biotechnology are closely based on European Union guidelines, the view of the European court would give the final decision, he said. Greenpeace's patent specialist, Christoph Then, admits Meier-Beck's ruling means that the case "is not running in our direction".

Brüstle says: "It's crazy that you are allowed to work on some hES cell lines in Germany and develop them for clinical purposes, but patenting your methods is deemed to be contrary to public order. Consultation with the European Court of Justice will hopefully contribute to harmonization of patent practice in Europe." ■

Alison Abbott

P. OOMEN/PHOTOGRAPHER'S CHOICE/GETTY

Biodiversity's bright spot

While species losses mount worldwide, conservationists in Brazil have made great strides towards saving the golden lion tamarin and its forest habitat from destruction. **Gene Russo** reports.

Biodiversity

At a farm less than two hours from the sprawl of Rio de Janeiro, a small group of ecotourists strolls to a patch of forest to see one of the rare victories in the fight to preserve Earth's dwindling biodiversity. Walking among trees near the town of Silva Jardim, the visitors are greeted with high-pitched squeals emanating from the canopy. A few seconds later, they spot tufts of brilliant orange fur flying from branch to branch as three families of golden lion tamarin (*Leontopithecus rosalia*) compete for territory in an area that should really accommodate only one.

The overcrowding of these small, twitchy primates in this forest fragment demonstrates both the successes and remaining challenges surrounding efforts to preserve the golden lion tamarin and its native habitat, the Atlantic forest of Brazil. Known as Mata Atlântica, these scattered fragments of forest dot the eastern coast of Brazil and contain one of the richest assortments of endemic species in the world, many of them endangered. And like the golden lion tamarin, the Atlantic forest almost disappeared because of development and an exploding human population. But both the flagship primate and its habitat have been saved, at least temporarily.

"When I come to the Atlantic forest, I think 'Thank God'," says Russ Mittermeier, president of Conservation International. Mittermeier, who travels the world visiting places in various states of ecological distress, lauds the area's conservationists, researchers and a frequently supportive public. "It's as good as it gets," he says.

By many measures, that's still not very good. The scraps of Atlantic forest occupy only about 10% of the area covered when Europeans arrived in 1500 (see map, opposite). And just 9% of those remaining bits are protected.

But the rate of deforestation has slowed in the past decade and in some places conservation efforts have even helped the forest to rebound. Beyond the hectares saved, the resurgence of the Atlantic forest shows how successful conservation can emerge through a combination of forces. The lessons learned here may resonate beyond Brazil, as nations seek to reverse the rapid loss of the planet's species. "The Atlantic forest shows that it's not a hopeless cause in these high-priority areas," says Mittermeier.

For this biome, it was the plight of the golden lion tamarin that helped motivate a conservation movement. As its habitat vanished, the population of these primates dropped to roughly 150 individuals in the 1970s, which caused conservationists and researchers to take aggressive action to save the species (see 'Bred to survive', overleaf). But it soon

became clear that the golden lion tamarin could not be preserved on its own.

Jim Dietz, a conservation biologist at the University of Maryland in College Park, remembers flying over the Atlantic forest in a helicopter in 1985 while there to investigate the tamarins. He was stunned by what he saw — a speckled landscape of forest fragments interspersed among pastures. The sight turned Dietz from a golden lion tamarin researcher to a forest conservationist. "I was down there to study mating systems," he says. "But I hadn't understood the grave nature of the threat."

Stemming the deforestation required a broad set of measures: new laws and governmental incentives, the commitment of researchers and conservationists, increased funding from international donors and the Brazilian government, and a growing community awareness. Lately, a boost has come from efforts to emphasize the forest's value as a source of water, a draw for ecotourism and a generator of other ecosystem services.

International pressure has also helped. Through the Convention on Biological Diversity, countries have committed to slow the rate of biodiversity loss and to protect 10% of their ecoregions by 2010. Although few nations will meet these goals, Brazil has set aside 16% of its land. Most of this is in the Amazon, but the biodiversity treaty has put pressure on Brazilian authorities to establish state parks in the Atlantic forest southwest of São Paulo, says Oliver Hillel, an officer in the convention's secretariat in Montreal, Canada.

"When I come to the Atlantic forest, I think 'Thank God'. It's as good as it gets." — Russ Mittermeier

In the balance

Preservation efforts have to fight against a long history of forest destruction in the Mata Atlântica. Shortly after landing in the region in 1500, Portuguese settlers began cutting down trees. By 1797, they had cleared so much land that Queen Maria the Pious of Portugal called for measures to stop the forest's destruction. In recent decades, sugarcane cultivation, logging and ranching have shattered the forest into fragments. Oil exploration on the coast and massive development — 70% of Brazil's 200 million people live within the Atlantic forest biome — have shrunk it further.

Estimates vary of how much forest survives. Measurements that count only fragments larger than 100 hectares indicate that just 7–8% of the forest is left. Calculations that include smaller fragments and more mountainous areas suggest that 11–16% remains. In either case, the forest is badly broken up. Roughly 83% of the fragments are less than 50 hectares in size (M. C. Ribeiro *et al. Biol. Conserv.* **142,** 1141–1153; 2009). The smaller the fragment, the harder it is to sustain existing ecosystems, especially those that include animals with large ranges, such as tamarin, jaguar, puma and many birds.

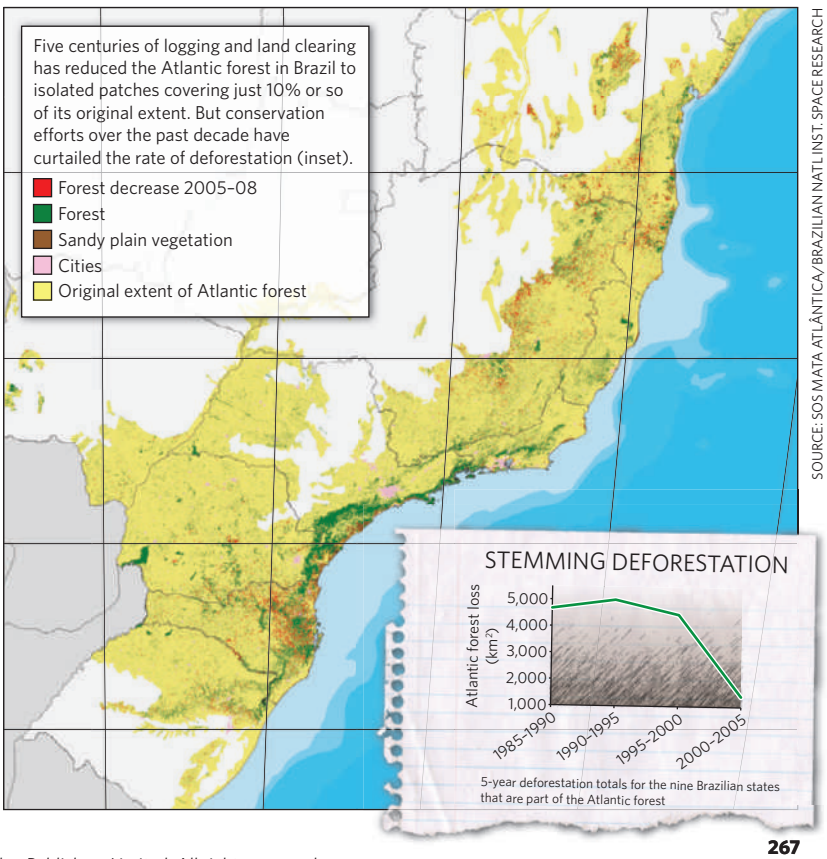
And yet, despite the devastation, these splintered ecosystems remain a hotbed of life. The forest, which stretches from the eastern tip of Uruguay northeast through Brazil's population centres to the tiny state of Rio Grande Do Norte, contains a far more varied range of elevations and climates than the vast Amazon basin. Its diverse environments support an unusually high number of species, many of them existing only in this biome. The Atlantic forest has an estimated 20,000 plant species, 8,000 of which are believed to be endemic. Roughly 190 animal species and 300 plants in the forest are considered threatened. Based on this density of life alone, there is a lot at stake.

The Brazilian government took a significant step to preserve its forests in 1965, when it revised the country's forest code. The law required landowners to preserve areas around rivers and forests on steep slopes. For the Atlantic forest, it dictated that 20% of any rural property must be maintained as a reserve.

More legislation in the 1980s and 1990s sought to conserve forest resources and ecosystems. But it was not until 2006 that Brazil passed a law specifically to protect the Atlantic forest, by demarcating its extent and its ecosystems and by requiring special permission for activities that damage them. The law prohibits the removal of vegetation in areas housing endangered species. It also provides for watershed protection, erosion control and the formation of corridors between forest remnants.

The long struggle to win even the promise of such safeguards was led in part by Marina Silva, a senator who had previously served as environment minister. She and others had battled an array of forces including the timber industry and landowners. Despite the environmental victories in the legislature, it has been difficult to win protection on the ground. "If the forestry code had been enforced and complied with, we would probably have 30% of the Atlantic forest left," says Lucio Bede, manager of Conservation International's Atlantic forest programme. Although the code specified punishment, it provided no real incentives for compliance. And enforcement has been lax, says Bede.

THE FRACTURED FOREST



SOURCE: SOS MATA ATLÂNTICA/BRAZILIAN NATL INST. SPACE RESEARCH

As for the Atlantic Forest Law, Silva says that too few municipalities are engaged in monitoring and recovery efforts. "The application of the law is not uniform throughout all Atlantic forest states," she says. The state of Santa Catarina, for example, has attempted to loosen the federal law with its own environmental code. But even with such problems, the recent law has helped raise awareness and slow deforestation, says Silva.

Notions of pristine environments, peaceful forest canopies and noble animals hold scant interest for poor people seeking jobs and income. So conservationists have recently tried a different tactic, stressing what intact forests can do for people living everywhere from rural towns to the megacities of São Paulo and Rio de Janeiro. These population centres rely on the forests to provide clean water, and deforestation threatens the watershed serving millions.

This approach relies on advertising the ecosystem services provided by the forest, a tactic adopted by conservation proponents and researchers in many parts of the world over the past decade. In Brazil, it is starting to yield results. "Municipalities are just beginning to understand conservation and environmental services," says Denise Marçal Rambaldi, secretary-general of the Golden Lion Tamarin Association and a key figure in the project to save the species. "Decisions made in a municipality are much more important for land use and planning than those in the federal government."

"We hope to create an economy based on ecosystem services."
— José Maria Cardoso da Silva

A river runs through it

In the city of Petrópolis, 70 kilometres outside of Rio de Janeiro, a small river runs near the summer residence of Pedro II, emperor of Brazil in the nineteenth century. The Piabanha river has become a focus of efforts to preserve the forest in and around the city. In 1992, Petrópolis and its surrounds became the first federally decreed environmental protection area. Now, engineers and local government officials are trying to conserve and enhance the vegetation along the river, in the hope that the Piabanha can serve as a thread that sews together forest fragments on the edges of the city.

The challenges are many. In spots, particularly at high altitudes, forest fires have eliminated native plants and allowed invasive grasses to take hold. Roads, condominium developments and shanty towns known as favelas impinge on the watershed, threatening water quality and encouraging erosion. According to Yara Valverde, a biologist and former head of the Petrópolis protection area, irresponsible development could lead to floods that would taint water sources. But the city imposes fines for environmental crimes. And guided by citizens and federal and local governments, the preservation efforts have allowed landowners to protect former grazing lands, allowing for some natural regeneration of the forest.

The progress, albeit incremental, in Petrópolis demonstrates the potential of emphasizing what the forest provides to the environment and community. "We hope to incorporate conservation planning into watershed planning and create an economy based on ecosystem services," says José Maria Cardoso da Silva, Conservation International's vice-president for South America. Currently, there are legal mechanisms to pay landowners to protect forests along a watershed; the crucial next step is to develop overall governance structures for each watershed across the Atlantic forest, says Cardoso da Silva. Government-sanctioned watershed committees have started to form that include local government, businesses and conservation organizations. By charging for water use, some committees have generated money for conservation, reforestation and sanitation management. The focus on ecosystem services, says Bede, makes clear how protecting the forest helps preserve the water supply.

Valuing ecosystem services, however, does not necessarily protect the forest's flora and fauna, because it shifts the emphasis away from saving species. Mittermeier endorses continued ecosystem-services projects but he warns: "If we

Bred to survive

Behind the animal enclosures at the Smithsonian National Zoo in Washington DC, Jennifer Mickelberg is greeted by a golden lion tamarin, which swings by and peers through the bars at a photo in her hands. The animal squeals at the glossy picture of a tamarin resting on a branch. "Yeah, that's your papa," coos Mickelberg.

She should know. Mickelberg is the assistant keeper of the 'stud book' — a database tracing the family tree of all 500 golden lion tamarins currently housed at 150 institutions worldwide and 3,500 of their ancestors. With these records and population-management software, the stud-book keeper decides which tamarins to breed together to maintain the captive population's genetic diversity.

The antecedents of this breeding programme began in the early 1960s, following a crash in the Brazilian population of wild golden lion tamarins. At first, scientists struggled to get tamarins to breed in captivity, but by the 1970s they had learned to include more protein in the animals' diet and to base social groups on monogamous pairs.

As captive breeding became more successful, tamarins from these programmes augmented the dwindling wild population in Brazil. At the same time, a network of dozens of zoos, led by the National Zoo, stepped up efforts to track genetic pedigrees,

J. DIETZ

hoping to minimize inbreeding. When researchers are considering a pair of potential mates, they examine mean kinship, the level of relatedness between an individual and a population, and another factor called the inbreeding coefficient, a measure of relatedness to a potential mate.

Currently the wild population of golden lion tamarins — found only in Brazil — has grown to the point at which animals raised in captivity are not being released. But the zoo population must remain healthy and genetically robust, in case a massive forest fire or some other catastrophe hits the wild population. As part of their work, researchers must also seek to keep artificial selection from happening in captivity. If the zoo environment were to encourage certain traits over many generations, there could be a propensity to produce animals adapted to captivity but maladapted to the wild. **G.R.**



FROM LEFT: J. RUDY; L. CLAUDIO MARIGO/NATUREPL.COM; A. MARTINS

Marina Silva (left) has helped protect the Atlantic forest, which borders cities such as São Paulo (centre). A parade in Silva Jardim shows local support for saving the golden lion tamarin.

focus on the ecosystem-service argument only, and don't focus on critical endangered species, then we could end up with good forest and good ecosystem services but no species."

Although different approaches could sometimes come into conflict, an emphasis on ecosystem services may actually dovetail well with protecting species, according to an unpublished study. Frank Wugt Larsen, a postdoc at Conservation International's Center for Applied Biodiversity Science in Arlington, Virginia, assessed 524 sites around the world in terms of the species they harbour and the services they provide, such as storing carbon in vegetation and providing clean water. He found that areas identified as the last-known habitat for a given species provide more ecosystem services than nearby 'control' areas.

In Brazil, the case of the golden lion tamarin shows how multiple conservation interests can converge to raise public awareness and move landowners to action. For more than 40 years, Marcos da Silva Freire and his family have owned 350 hectares, outside Rio near Poço das Antas, the first federal biological reserve in Brazil. Freire, an immunologist, works in Rio like many middle-class landowners, but he maintains the land and visits when he can. In the late 1980s, researchers contacted him and his father about introducing two groups of tamarins, bred in zoos, onto their land. They agreed, and theirs was the first of several farms to offer the animals indefinite accommodation.

But interspecies altruism is not the only factor that motivates Freire. He enjoys the tamarins — and he has considered establishing his own ecotourism venture to bring in additional revenue.

Other conservation efforts could help sustain poor farmers and, at the same time, enlist them to protect the forest. On a hot, clear day in August, Adeildo Ataliba proudly shows visitors his plot of land near the Poço das Antas reserve, one of the main homes of the golden lion tamarins. Walking through tall trees and shrubs, he points to plants he has recently sown, including seedlings of coffee, yucca, guava and jatoba. Nearby is a large banana tree and a single stalk of sugarcane.

"We have ten years to ensure legal protection and sustainability."
— Denise Marçal Rambaldi

Just a few years ago, this land was a field of ankle-high scrub. As part of an agrarian resettlement effort, Ataliba and his family were one of several landless families who received a plot from the government. But many of these families struggled because they lacked agricultural training and an environment conducive for traditional cash crops. A few of the farmers even resorted to hunting in the nearby reserve. In an effort to aid the farmers and protect the reserve, the Golden Lion Tamarin Association — whose mission has expanded beyond tamarins since it was founded in 1992 — established 'agroforestry systems' of native tree species planted together with vegetables and fruits. This helps restore forest corridors for wild animals and yields produce for the families. The association helps out farmers by providing seedlings and technical assistance.

Conservation groups have even more ambitious aims for the future. A coalition of non-governmental organizations, research institutions and private companies hopes to double the amount of Atlantic forest by 2050. Some of the money for this effort, expected to cost tens of billions of dollars, could come from payments to store forest carbon, an ecosystem service that could become much more valuable under a new climate treaty.

As the various stakeholders in Brazil explore different approaches to preserving the Atlantic forest, the golden lion tamarins continue to multiply in their newfound territory. Currently, their population in the wild has swelled to around 1,500 on more than 10,000 hectares of protected land. But in many places, pastures and roads prevent the growing families from expanding into new territory. Dietz is seeking funding to join up enough isolated tamarin populations via forest corridors to enable gene flow throughout the entire population. "It would be self-sustaining in perpetuity," he says.

Rambaldi warns that the tamarins and their habitat still face significant threats. "We have ten years to ensure the legal protection and the sustainability of these fragments," she says. Otherwise, urban pressures and oil exploration threaten to claim land. And the fragmented forest, says Rambaldi, will become just an empty collection of trees. ■

Gene Russo is editor of *Naturejobs*.

See Editorial, page 251, and online at www.nature.com/darwin.

PUTTING A PRICE ON NATURE

Gretchen Daily knows the value of ecosystems — but can ascribing financial worth to them help to maintain biodiversity? **Emma Marris** meets an ecosystem-services evangelist.

When Gretchen Daily was young, she watched acid rain slowly killing the forests around her in Germany's Taunus mountains. As a researcher, she cut her teeth studying extinctions under Paul Ehrlich, an ecologist famed for his predictions of mass starvation. Last month, Daily travelled to Japan in advance of next year's meeting of the Convention on Biodiversity in Nagoya, where the world will hear how spectacularly that treaty has failed to protect the planet's species. All this makes it remarkable that Daily is, eternally, sunny.

Daily's enthusiasm bubbles forth in meeting rooms around the globe as she promotes the 'ecosystem-services' approach to conservation, of which she has become the world's most passionate proponent. Her argument — and the argument of a group of like-minded researchers — is that undeveloped nature provides services to human society such as clean water and flood protection that can be valued in financial terms that are large enough to justify protecting it (see 'Ready to serve'). She also believes that this protection can be achieved by installing sufficient financial incentives to make owners want to preserve this bounty.

In 1997, Daily edited an influential book that made a first coherent case for saving the planet with cash (*Nature's Services: Societal Dependence on Natural Ecosystems*). That same year, a much-discussed paper estimated the total worth of 17 of Earth's major ecosystem services at US$33 trillion a year (R. Costanza *et al. Nature* **387,** 253–260; 1997). The resultant buzz propelled the idea into the Millennium Ecosystem Assessment report of 2005, which used ecosystem services as a framework to discuss the state of the planet and how to preserve it.

Daily, now working at Stanford University in Palo Alto, California, is gearing up for a major publicity push for the concept in 2010, the International Year of Biodiversity. But this increased attention could also highlight the flaws of the ecosystems-services approach, one of which is its uncertain ability to protect biodiversity: in some cases a biodiverse ecosystem does not necessarily provide services that are more financially valuable. Not that these arguments will stop Daily. "Gretchen is going a zillion miles an hour and she's got this crusade, if you will," says Steve Polasky, an environmental economist at the University of Minnesota, St Paul. "You often get these crusaders and it is all about them — but with Gretchen it is really about getting ecosystem services on the agenda."

Economists have been working on attaching monetary value to components of natural systems since at least the 1960s, evaluating the cost of damage caused by oil spills, for example. But environmental activists and conservationists didn't pay this work much attention. Many felt that nature should be saved not for it price, but for its own sake.

Daily's conversion happened gradually. Born in the United States, she spent her adolescence in Germany in the midst of early 1980s environmental protests. "It was amazing to see the demonstrations out in the street protesting acid rain and everything connected to it," she says. The experience convinced her of the value of using science and activism to tackle environmental problems.

She did both, working at the Worldwatch Institute, an environmental think tank in Washington DC, as an undergraduate in the mid 1980s and then applying for her graduate studies to work with Ehrlich at Stanford. At a field station in Gothic, Colorado, in the early 1990s, Daily mixed with Ehrlich's influential friends. These included Peter Bing, a rich businessman and then chairman of the board at Stanford, who "knocked some sense into me" on day-long hikes, says Daily, encouraging her to talk with business people in their language — economics — rather than see them as the enemy. She also met Tim Wirth, who was then one of Colorado's senators and an early advocate of cap-and-trade approaches to combating pollution.

Daily became convinced that such incentive schemes were the way to save the environment.

"I work day and night. I'm basically a fanatic."
— Gretchen Daily

Ready to serve

A selection of nature's 'services':
Provisioning: timber, fish, wild game, fruit and fungus, even moss and foliage for floral arrangements.
Regulating: water filtration and capture, flood protection, carbon sequestration.
Cultural: recreation, education, aesthetic and spiritual contemplation.

She won financial support from various foundations to prepare, edit and publish *Nature's Services* and she has hardly looked back since. "There has been tremendous behind-the-scenes progress," says Daily. The concept has been widely embraced by policy-makers. Ecosystem-services projects are now so thick on the ground that one needs a dictionary to keep track of all the acronyms. Among those that Daily salts her conversation with are TEEB (the Economics of Ecosystems and Biodiversity), a European study on how much money the continent might be losing through ecosystem loss, and IPBES, the Intergovernmental Platform on Biodiversity and Ecosystem Services, a proposed scientific advice generator modelled on the Intergovernmental Panel on Climate Change (IPCC).

Seeing solutions

When Daily is not spreading the word on ecosystem services to scientists, policy-makers and broader audiences, she is pursuing her own studies in places such as Costa Rica and Hawaii. In Hawaii, Daily brought together parties who had long fought over land use — ranchers, native Hawaiians and water and power companies — and persuaded them to write a report, now under review, to the state legislature that recommended reforesting areas of ranchland. Daily got them all to sign up to the same recommendations in part by focusing on their common concerns about land being converted to be used for high-end homes. Under the new proposal, the ranchers would be paid for reforesting, the native Hawaiians would have access to the forest and the trees would retain rainwater and keep salination of the drinking water supply at bay. "It was really stunningly easy to get people together in dialogue," she says. Long-time collaborator Peter Kareiva, chief scientist of the Nature Conservancy in Seattle, Washington, says of Daily that "people around her are energized. She's built relationships. She just has that personality. She sees a solution."

Daily focuses much of her energy on the Natural Capital Project, a joint effort she brokered between Stanford, the conservation group WWF and the Nature Conservancy. The project, which she co-directs, is developing a software system to help people weigh up the value of land in terms of ecosystem services,

Gretchen Daily hopes that ecosystem services can buy time for biodiversity.

V. EVANS

alongside its value for building houses or other development. Maps of an area are layered with information — much of which can be displayed in dollars — such as which parts of the landscape are best for filtering water, where the real-estate is most valuable, where the most carbon can be stored and where the biodiversity is highest. Such maps could help governmental organizations evaluate, for example, the cost of building on land that provides free water filtration if the development would then require the construction of a costly water plant. "We can bring about this transformation if we supply tools that make it easy for decision-makers to compare alternative scenarios," she says.

Over the past few years, the maps have been used by officials in China's upper Yangtze River basin to help plan urban and agricultural expansion and dam construction. "I don't earn that much money, but I am laying down my life for all this," Daily says. "I work day and night. I am basically a fanatic."

The idea of ecosystem services has its critics. John Echeverria, an environmental lawyer at Vermont Law School in South Royalton, says that paying landowners not to damage the environment sets up an expectation of reward for refraining from bad behaviour, and a financial obligation for future taxpayers. "The implicit message of agreeing to pay is that they should be entitled to proceed to destroy nature," says Echeverria. Instead, he suggests, landowners should in general be expected to do the right thing and be punished when they don't — the model enforced by the US Endangered Species Act and equivalent legislation in other countries.

Daily contends that the Endangered Species Act and similar laws are failures because their restrictions and penalties have angered many landowners. They also create an incentive for landowners to remove any endangered species from their land before the authorities find out about them. This approach has "led to these past decades of lose–lose battles on the environment", she says.

Richard Carson, an economist at the University of California, San Diego, is a fan of Daily's work, but he says that her pitch tends to focus on the easy cases. "If there is a problem, it is that she has created the impression in people that if you just think about these things in the right way, everyone is going to come out ahead." Daily agrees that she is going after the win–win situations, and says it is because there are still so many easy gains to be made. But eventually, she knows, there will be some tough decisions. If a fish species is close to extinction, it may be necessary to completely close the fishery for some years to ensure that the service (provision of fish) is maintained in the future; and it is difficult to make that decision a 'win' for fishermen.

Rationale for destruction?

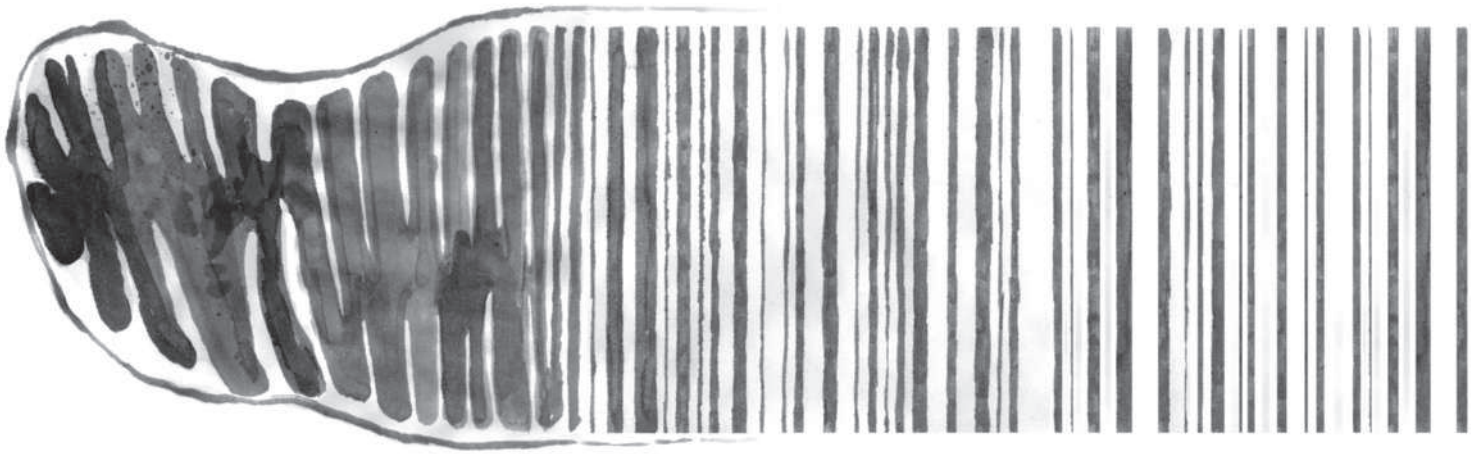
There is another fundamental limitation to the ecosystem-services framework: some services provided by an ecosystem are simply not considered valuable enough to warrant protecting. Biodiversity is particularly problematic. In some cases, a monotonous plain of non-native grass delivers better and cheaper ecosystem services, measured in water filtration, carbon sequestration and flood protection, than a diverse marsh. Attaching explicit values to things can provide a rational basis for ignoring them.

Nevertheless, Daily says, the ecosystem-services approach can save many places with high biodiversity — and at the very least it will give certain ecosystems time until society shows more willingness to protect them for other reasons. "I think it is going to be a long haul for biodiversity for its own sake. For me, ecosystem services is a strategy to buy time as well as getting buy-in." Such sentiment reveals that the ecosystem-services approach is not necessarily that different from conventional environmentalism. Advocates of both viewpoints believe that nature is intrinsically valuable, and they hope to preserve nature by appealing to this belief in others or, where it is absent, by creating it. The difference is that Daily works to convince others by showing them the profitable side of nature first.

Peering through the blur of her hectic work life — the conferences, authoring, media interviews and research — it is clear that Daily isn't just a sunny personality. She is a true optimist. She believes that people can and will save the planet's biodiversity — not just because there is something in it for them, but because, eventually, they will care. ■

Emma Marris writes for *Nature* from Columbia, Missouri.

See Editorial, page 251, Opinion, page 277, and the biodiversity special at www.nature.com/darwin.



On the origin of bar codes

Genetic sequences in a cell's mitochondria can be used to accurately determine species. Could this be because they are responsible for creating what they identify? **Nick Lane** investigates.

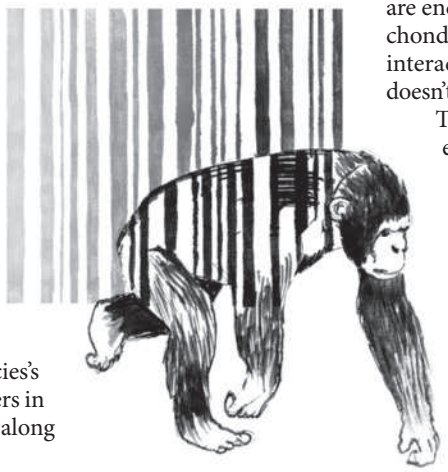
Mitochondria, the cell's energy producers, keep a low profile in terms of their genome. Descended from free-living bacteria that took up residence within other cells some 2 billion years ago, they've maintained a modest genetic repertoire — a mere 37 genes in vertebrates, compared with more than 20,000 in a nucleus.

Yet within this little genome, researchers have pinpointed a 648-nucleotide stretch as the ultimate identifier of species, dubbed the DNA bar code. The sequence can distinguish between closely related species such as humans and chimps and even classify new species from identical-looking ones, such as the blue-flasher butterfly (*Astraptes fulgerator*), which has since been divided into ten separate species, verified by the habitats, lifestyles and diets of their caterpillars.

The DNA bar code has been both praised and attacked for its simplicity. Many assume that it misses taxonomic subtleties that can be revealed only through traditional systematics or more extensive sequencing. However, proponents take the criticisms in their stride. "The fact is these short sequences yield surprisingly accurate information about the composition of the entire genome," says Donal Hickey, an evolutionary biologist at Concordia University in Montreal, Canada.

A part of a mitochondrial gene was chosen simply because it worked better than other sequences, and researchers assumed that the sequences became unique after two species split from their common ancestor. But what if DNA bar codes work for a deeper reason? Molecular biologist Dan Mishmar and his colleagues at the Ben-Gurion University of the Negev in Beer-Sheva, Israel, have been collecting evidence to support a new hypothesis: that mitochondrial sequences — such as those that give rise to each species's unique bar code — might actually be powerful drivers in the process of speciation[1]. Rather than simply going along for the ride, they say, they could be responsible for undermining the reproductive compatibility within a species when they conflict with sequences in the nucleus.

Genes in the nucleus must adapt quickly to partners in the mitochondria.



Although it is largely — some would say almost entirely — speculative, some evidence supports the idea, including data from a project called the Barcode of Life Initiative, says founder Paul Hebert from the University of Guelph in Canada. The initiative is a collection of research projects, organizations and individuals devoted to developing DNA bar-coding as a global standard for identifying species. So far, it has described the bar codes of almost 65,000 species. If true, Mishmar's hypothesis could connect the origins of species with biodiversity in a satisfying way. "Reproductive isolation may arise through a complex genomic ballet — a *pas de deux* between the mitochondrial and nuclear genomes," proposes Hebert.

Mitochondria matter

Cell survival depends on respiration, which takes place in the mitochondria through a series of large protein complexes, each built from as many as 43 subunits. These subunits are encoded by genes that reside in both the mitochondrial and the nuclear genome, and they must interact intimately with each other or respiration doesn't work.

Take the enzyme cytochrome oxidase, for example, which handles the final step of cell respiration. In mammals, the complex is composed of 13 subunits, three of which — including subunit 1, the bar-code gene — are encoded by mitochondrial DNA, and ten by nuclear genes. If the subunits of cytochrome oxidase don't work together properly, electrons are not passed to oxygen and respiration fails, triggering the death of the cell.

Making this cooperation even trickier is that fact that the two genomes evolve in dif-

ILLUSTRATIONS BY CHRISTINE BERRIE

ferent ways. Nuclear genes are mixed by sexual reproduction, every generation different alleles are introduced, whereas the mitochondria divide in a simple asexual fashion. Moreover, the mitochondrial gene sequences generally change much faster from generation to generation than the nuclear ones — typically 10–30 times faster.

Given the penalty for failure, it is hardly surprising that the two genomes have adapted to work together. Every generation the mitochondrial genes are tested against the new nuclear background of the offspring. If they don't work, there can be a developmental failure or a serious reduction in fitness after birth, referred to as hybrid breakdown. The outcome is that selection acts to ensure that the two genomes function properly together; and there is plenty of evidence showing that, despite their different modes of evolution, changes in one genome bring about a strong selection for compensatory changes in the other[2].

In 2006, Mishmar found evidence for this co-adaptation while working with Doug Wallace, a geneticist at the University of California, Irvine. They found that the primate nuclear genes that encoded mitochondrial proteins had a similar number of changes to the mitochondrial genes and that these genes evolved ten times faster than other genes in the nucleus[3]. In other words, the mitochondrial and nuclear genes adapt to each other within a population, and the process must happen quickly because the mutation rate is so high in mitochondrial DNA.

In copepods, a mitochondrial gene mismatch makes hybrids less fit.

Mishmar started thinking about the implications of this. If one set of genes with a critical effect on survival evolves ten times faster than other genes, two populations of the same species could quickly diverge, possibly even driving a wedge between them reproductively.

Where's the evidence?

Perhaps the best evidence for reproductive incompatibility comes from marine biologist Ron Burton and his colleagues at the Scripps Institution of Oceanography in La Jolla, California. They have shown that incompatibilities between mitochondrial and nuclear genes can seriously undermine the fitness and fertility of the intertidal copepod *Tigriopus californicus*[4] — a crustacean whose small size and abundance often earns it the name 'insect of the sea'.

Burton cross-bred individuals from nearby populations of *T. californicus* that don't usually interbreed so that the mitochondria from one population ended up paired with the nuclear genes of the other — a process known as introgression. The mismatch suppressed reproduction and cellular respiration of offspring by as much as 40%. Low respiration was also linked to slow juvenile development and poor survival — altogether, a serious reduction in fitness.

If, in effect, all these hybrid offspring are runts with low fertility, then they are less likely to survive and reproduce than the offspring of pairings within the same population. Over time, what started as partial reproductive incompatibility ends up as a total failure for the two populations to produce viable offspring when crossbreeding — which is to say, speciation.

Burton, however, is careful not to over-claim. And although others praise his work, they too caution against over-interpretation. "Just because a genetic interaction causes hybrid problems doesn't mean it was responsible for speciation in the first place," says Jerry Coyne, an evolutionary biologist at the University of Chicago in Illinois. At least 200 genetic interactions are known to cause inviability when *Drosophila simulans* is crossed with *Drosophila melanogaster*. But when the common ancestor originally split into these two species, only one of them was probably the driver, because any one alone causes inviability. Other differences would only have emerged after the species had diverged.

Pinning down exactly which of these 200 interactions is responsible for speciation is an arduous task, and Coyne has seen no hint that mitochondrial interactions have a role. Far from it — he believes that the existing evidence refutes the postulated role of mitochondria in speciation.

"Closely related species living together often have identical or very similar mitochondrial DNA, even when their nuclear DNA is more diverged," Coyne says. Mitochondrial genes seem to flow between closely related species more easily than nuclear genes do, for unknown reasons — although this does not happen frequently enough to totally derail bar codes.

David Rand, a molecular evolutionist at Brown University in Providence, Rhode Island, has come across similar problems. With Brown's Colin Meiklejohn and Kristi Montooth, now at Indiana University in Bloomington, Rand has transplanted mitochondrial DNA from different species of fruitfly into the nuclear background of *D. melanogaster*, and found little evidence of the hybrid breakdown that would support the idea of mitochondria controlling speciation.

"I'm sure there are cases in which mitochondrial–nuclear interactions have had roles in speciation" says Rand, "but how common is it? Few people have dissected this carefully enough yet."

As one of the few who has approached the topic, Burton is not surprised that examples are scarce as yet. The degree of hybrid breakdown, he says, depends on how fast the mitochondrial DNA mutates — and that is under tight genetic control, and varies from species to species. Fruitflies and other heavily studied model systems don't have particularly high rates of mutation, he says. "That makes them the last place to look for evidence."

The best places to look, according to Burton, are taxa that have high mitochondrial DNA mutation rates, which range widely from rodents to Galapagos tortoises, from snails to copepods — even yeast. Populations from these taxa are much more likely to break down if crossed with other populations of the same species, as a first step towards reproductive isolation and speciation — as happens in Burton's copepods, and in yeast[5].

But why do some organisms have a much faster mutation rate than others? The mitochondrial DNA of birds, for example, changes at a quarter of the speed it does in mammals[6]. So what is the benefit of a fast mutation rate; and what, if anything, does it say about speciation?

Wallace says the answer is simple: rapid mitochondrial

mutation is adaptive. "Reproductive success requires adapting to different food sources — carbohydrates, proteins or fats — and climates from icy cold to intense heat or humidity. A lot of this adaptation goes on in the mitochondria," he says.

A fast mutation rate can quickly produce variants that are suited to the changing environmental conditions. The only drawback is that a high rate of change should also lead to lots of negative mutations, and ultimately a meltdown for some populations. It's hard, on the face of it, to see how that could be adaptive.

Last year, however, Wallace's group came up with an answer to this conundrum[7]. In mice, severe mitochondrial mutations are eliminated in the germ line — eggs with mitochondrial mutations fail to develop, meaning that the eggs that do develop are more likely to be healthy. The fast mutation rate generates variation coupled to a developmental filter that gets rid of the most detrimental mutations before they have the chance to undermine the health of an animal's offspring.

This means that a high mutation rate in mitochondrial DNA can be adaptive. It is beneficial and can be selected for. But equally, it affects speciation. A fast mutation rate means that the genes controlling respiration change quickly over generations. And that, in turn, increases the chance of mismatch between mitochondrial and nuclear gene sequences if and when individuals outbreed with other populations.

This scenario paints a radically new picture of mitochondrial genes as being tightly regulated by selection. Until recently, most had thought of them as little more than 'neutral markers'.

Evolutionary biologist Nicolas Galtier and his colleagues at the University of Montpellier 2 in France have been chipping away at this neutral view, and have found that in most species, the variation in mitochondrial DNA is surprisingly restricted[8]. If mitochondrial DNA really is a neutral marker, mutations should build up quickly over time, giving plenty of variation. But if mitochondrial DNA is subject to periodic bouts of selection then much of this variation would get purged.

That's what Galtier sees in the data. Each such 'selective sweep' wipes out the common ground between species, and leaves little variation within a species — a nifty trick if you're looking for a bar code.

From bar codes to species

By definition, a DNA bar code is a unique identifier. Between any two humans, it varies at no more than two positions[2]. By comparison, humans differ from chimpanzees at approximately 60 sites, and from gorillas at about 70.

Much the same is true of the other 65,000 or so species that have had their bar codes sequenced. Just as Galtier has found in mitochondrial DNA in general, DNA bar codes vary remarkably little within a species, but have little or no overlap between species. Given the characteristic mitochondrial combination of rapid mutation and limited variation, only two processes can generate such a pattern: natural selection, or genetic drift after a population 'bottleneck'.

The human genome seems to be a result of the latter. As mitochondria are inherited only through egg cells, human mitochondrial DNA can be traced back to a shared common female ancestor, thought to have lived in Africa 170,000 years ago, and named Mitochondrial Eve. But Mitochondrial Eve did not live alone, so why has all humanity inherited her mitochondrial DNA? The standard answer invokes a bottleneck that reduced the number of humans down to a few thousand individuals, whose descendants took over the world. From a limited repertoire of mitochondrial DNA, one type happened to become fixed through genetic drift — no selection necessary.

Mitochondrial DNA's high mutation rate may help some organisms adapt to changing environments.

Like Galtier's findings, data from the Barcode of Life Initiative raise doubts about this interpretation. All species show the same lack of bar-code diversity. Although it is easy to imagine that humans passed through a bottleneck 170,000 years ago, it's hard to believe that exactly the same thing happened in all species. "Did herrings really pass through an equally recent population bottleneck? Anchovies too?" asks Hebert. In his view, the only explanation is heavy selection across whole populations.

That makes a lot of sense in the context of Wallace's ideas on mitochondrial adaptation to climate or food. Mitochondrial genes mutate rapidly, generating variation that is subject to selection whenever environmental conditions change. That's suggestive of a selective sweep. And as Mishmar has been noticing, mitochondrial sequences place a premium on compatible nuclear genes, forcing fast changes on them. If individuals from nearby populations are then mated, the outcome is hybrid breakdown — the serious loss of fitness chronicled in Burton's crossed copepods. The rate at which any of this happens depends on the mitochondrial mutation rate — fast rates leads to fast divergence and a greater likelihood that mitochondria will have a role in speciation. In this view, the DNA bar code does not merely track species — it could very well create them.

Is this really a powerful driver of speciation? It's still too early to say. Mishmar is about to embark on a series of studies on cellular respiratory function in mismatched populations, and he challenges others to follow. If bar codes are fundamental, then the best place to look will be in closely related species with distinct bar codes. Mismatches in nuclear and mitochondrial genes between such populations should lead to hybrid breakdown that can be rescued simply by backcrossing. There is already a database of 65,000 species out there — and everything to play for. ■

Nick Lane is the first Provost's Venture Research Fellow at University College London and author of *Life Ascending: The Ten Great Inventions of Evolution*.

1. Gershoni, M., Templeton, A. R. & Mishmar, D. *Bioessays* **31,** 642–650 (2009).
2. Blier, P. U., Dufresne, F. & Burton, R. S. *Trends Genet.* **17,** 400–406 (2001).
3. Mishmar, D. *et al. Gene* **378,** 11–18 (2006).
4. Burton, R. S., Ellison, C. K. & Harrison, J. S. *Am. Nat.* **168,** S14–S24 (2006).
5. Lee, H. Y. *et al. Cell* **135,** 1065–1073 (2008).
6. Nabholz, B., Glémin, S. & Galtier, N. *BMC Evol. Biol.* **9,** 54 (2009).
7. Fan, W. *et al. Science* **319,** 958–962 (2008).
8. Bazin, E., Glémin, S. & Galtier, N. *Science* **312,** 570–572 (2006).

See Editorial, page 251, and the whole biodiversity special at www.nature.com/darwin.

Experts and democracy

Specialist advice can be invaluable in shaping policy, but, argues **Colin Macilwain**, democracies need to keep a careful eye on the powers acquired by an unelected elite.

From the perspective of political leaders, scientific advice is never a panacea. At its best, it can offer views and ideas that will inform sensible and resilient public policy. But it can also impart information that is unwelcome to politicians and senior civil servants, and if that disconnect goes public, the advice and the adviser can become more trouble than they're worth. That's one reason why the planned appointment of the first scientific adviser to the European Union (EU), announced in September, has been less than universally welcomed by senior officials in Brussels.

One such disconnect flared up in Britain late last month, and resulted in the summary dismissal by the home secretary, Alan Johnson, of David Nutt, the chair of the UK Advisory Council on the Misuse of Drugs.

The relationship between advisers and government is distinct to each society. In the United States, for example, the party affiliation of most citizens, including scientists, is in the public domain. Some staff advisers, as well as members of advisory panels on contentious topics, are selected, in part, on party political grounds.

In some European nations, including Britain, senior scientists manage to stay above the party-political fray. But this approach carries its own hazards. One is that there is less circulation of advisers when parties move in and out of power. Another is the conceit that these advisers have no political views of their own, and speak only for science.

Yet the delivery of scientific advice on contentious policy matters is inherently political. When Robert May or David King, as chief scientific advisers to then Prime Minister Tony Blair, vocally supported genetically modified crops and nuclear power, they were taking political positions — on which May, King and Blair were all in happy agreement.

When such advisers stay silent on the British government's largest single technical project — replacing the Trident nuclear submarine fleet — that is a political act. And when Nutt, having survived a run-in with a previous home secretary on drugs policy, went on the radio show *Today* to air his views on drugs policy and announced that he'd be airing them again at a lecture in London, that too was a political act.

Nutt's firing has triggered a vigorous debate on what scientific advice is and what it is for.

WORLD VIEW

But some scientists' contributions to this discussion have pointed to hubris on the part of a scientific establishment that has, over 12 years of Labour government, moved steadily closer to the levers of power.

The idea that the influence of scientific and other elites can become malign is not new. It is almost 50 years since US President Dwight D. Eisenhower planned to warn of the perils of a 'military–industrial–scientific complex', before his science adviser, James Killian, persuaded him to remove the word 'scientific' from his delivered farewell address. Some of Eisenhower's concerns were specific to his time and place. But he was also highlighting the fact that specialist advisers can exercise a large and seldom-noted influence on government — an influence that comes at the expense of other groups in society.

In wresting Labour from its old ideological moorings, Blair and his successor Gordon Brown chose to rely heavily on 'experts' from science and, until last year, finance. These experts spoke New Labour's language. They weren't bolshie ecologists or sociologists from the provinces. Rather, they were upstanding (almost always) men of Oxford, Cambridge or Imperial College, who supported free markets, genetically modified crops, nuclear power and the independent nuclear deterrent. All on purely scientific grounds, of course.

For the scientists, this partnership has worked well. Money has poured in to universities, with the science budget doubling since 1997 in real terms to £6 billion (US$10 billion). More than 70 scientific advisory panels, involving hundreds of scientists, have exercised a growing influence on national policy.

This partnership has offered the nation several advantages, including the revival of the universities and a more-informed public discourse on issues such as global warming. But the benefits to Labour have been less clear-cut. The money invested in science has not triggered the industrial modernization so fervently sought by Brown. And the government's marriage to scientific and financial expertise has coincided with its divorce from Labour's traditional base of support. The rise of the 'expertariat' has not caused the sharp fall in voter turnout, the tidal wave of public cynicism regarding politics and the sense of public life in crisis. Nonetheless, the government's heavy reliance on what looks from the outside like a narrow circle of advisers echoes the situation contemplated by Eisenhower.

As Nutt repeated his view that cannabis and some other recreational drugs are less harmful than alcohol or tobacco, Labour was fighting a difficult by-election in Glasgow North East — one of the most drug-ravaged and economically deprived locales in western Europe, where it can barely afford to sound 'soft on drugs'. Johnson's admittedly rash firing reflected his frustration that, in the run-up to a general election, public pronouncements from persons closely associated with the government cannot conflict too brazenly with the government's own position.

A 6 November statement, signed by 28 scientific leaders, reiterated current best advisory practice — including the right of advisers to speak freely in a personal capacity, and of committees to operate without political interference — and sensibly called on the government to reaffirm its policy on this. More emphatic demands, however — such as calls for scientific panels to set policy or to be protected from dismissal by elected officials — are dangerously naive.

The EU, in developing its advisory apparatus to match the expansion of its political authority under the newly ratified Treaty of Lisbon, can learn from this unseemly spat. Like any polity, the EU will benefit from listening to scientists. But the EU is already afflicted with a reputation for remoteness from the people. It cannot afford to leave the impression, as New Labour has done in Britain, that it is relying on an elite for advice. Vital expertise resides in many segments of society — including ones that consider themselves relatively marginalized, such as social workers or even engineers. In a democracy, the advice of scientists alone cannot be relied upon to deliver sound public policy. ■

Colin Macilwain is based in the United Kingdom, and his column will appear in the third issue of each month.
e-mail: cfmworldview@gmail.com
See go.nature.com/ILx8PC for more columns.

CORRESPONDENCE

Boreal forests' carbon stores need better management

In the run-up to next month's climate-change treaty negotiations in Copenhagen, there is a pressing need to inform policy discussions about the importance of carbon management of northern boreal forests, as well as of tropical forests.

Boreal carbon pools account for more of the overall carbon stock than tropical forests — a minimum of 559–703 gigatonnes, compared with 375–428 gigatonnes — and store twice as much carbon per unit area (see R. T. Watson *et al. IPCC Special Report: Land-Use Change and Forestry* Cambridge Univ. Press; 2007, and E. S. Kasischke *Fire, Climate Change, and Carbon Cycling in the Boreal Forest* Springer; 2000).

In tropical forests, carbon flux is equilibrated between sequestration in growing trees and loss from decay of dead trees. Boreal ecosystems, on the other hand, accumulate carbon over millennia in soils, peat and sediments and under permafrost, because low temperatures prevent biotic breakdown and release of accumulated carbon.

The large carbon stocks and sequestration potential of tropical and boreal regions are under threat from deforestation and habitat degradation. The rapidly expanding human industrial footprint in boreal regions in Canada and Russia, for example, will increase the risk of releasing emissions from the vast carbon stores of these areas. To reduce climate disruption, efforts are needed at international, national and regional levels to develop incentives for encouraging protection of these intact ecosystems.

Stuart Pimm Nicholas School for the Environment and Earth Sciences, Duke University, Durham, North Carolina 27708, USA
e-mail: StuartPimm@me.com
Nigel Roulet Department of Geography and the Environment, McGill University, Montreal, Quebec H3A2K6, Canada
Andrew Weaver School of Earth and Ocean Science, PO Box 3065 University of Victoria, Victoria, British Columbia V8W 3V6, Canada

Legal and practical pitfalls in making use of patents

Authors of research papers should use and cite online patent databases more frequently, according to a recent Correspondence (*Nature* **461,** 340; 2009), but from a US perspective, this is unsound advice.

An employee publishing a patent citation may be exposing his or her employer to liability for triple damages, and many firms ask their technologists to remain ignorant of the patent literature. US patent law awards triple damages for the period in which an infringer wilfully violates intellectual property. Ignorance is a valid defence against such a claim, so, before contemplating legal action, owners of intellectual property will often alert potential infringers, preemptively starting a wilful-violation clock. Widespread patent citation could even lead corporate lawyers to advise corporate scientists to avoid even reading or citing papers that cite patents.

Although patents represent rigorously reviewed novel work, they cannot be compared to peer-reviewed academic articles. The patent defines the invention (in its claims) and argues for its priority and novelty, minimizing the relevance of prior publications, whether legal or academic, along the way.

An examiner issues a patent only after rigorous review, but his or her report focuses on the patent's claims. Therefore the scientific validity of the patent's arguments and supporting data are not necessarily central to this review.

If the inventors (or others) later recognize serious flaws in the patent's data or scientific reasoning, they are under no obligation to retract or correct the patent. An invention is still valid and enforceable, and remains in public databases, even when the inventor got the science wrong.

David Piehler Fields and Waves, 17500 Cabrillo Highway South, Half Moon Bay, California 94019-8533, USA
e-mail: piehler@fields-and-waves.com

Water should take centre stage at climate talks

Political agreement on water rights and usage should be at the heart of climate-change conference discussions (see www.nature.com/roadtocopenhagen). Policy-makers, whose real agenda may be their standing with their electorates, should recognize that water is an ideal vehicle for reaching a useful aggregate agreement.

Water is where the impact of human-induced climate change is felt most keenly at a regional level. Demand is increasing, and the economic and social costs of extreme events such as floods and droughts are high. We should be ready to invest more in agricultural and industrial water-use efficiency, as well as in regional adaptation measures — bearing in mind that both water transport and desalination are very energy-intensive processes.

Globally, more than 32 trillion litres of treated water leak from urban water supply systems every year, according to the energy and water department of the World Bank (see go.nature.com/fLFTId). Last year, the UK water-industry regulator Ofwat reported that 22% of the water supply for England and Wales leaked away — 3.3 billion litres a day (see go.nature.com/3hJC7f). In developing countries, 40–50% of the supply may be lost this way.

The World Bank expects adaptation to climate change to cost US$75–100 billion a year until 2050 in developing countries alone (see go.nature.com/JCqkYq). The question of how to finance this is on the Copenhagen agenda.

One of the key requirements for success is to link incentives for sustainable water management with adaptation and mitigation strategies at a local scale. Local voters and governments are concerned with economic development, which is strongly linked to the security of water supplies, and are therefore most likely to support such measures.

Yulia Timoshkina Judge Business School, University of Cambridge, Cambridge CB2 1QA, UK
e-mail: it222@cam.ac.uk

Sensible measures to guard India's groundwater supply

Satellite-based estimates of groundwater depletion in northwestern India reported by Matthew Rodell and colleagues (*Nature* **460,** 999–1002; 2009) should be backed up with precise ground-based information for the whole of India, taking into account regional variations in rock types, aquifers and watersheds. This will help to ensure effective local remedial action.

I have worked among poor farmers in Indian villages and believe that, if the government withdraws electricity subsidies for operating groundwater pumps in northern India, it could lead to drought in agricultural regions. Suicides among farmers have already occurred in response to drought in Maharashtra and Andhra Pradesh.

As the authors point out, use of groundwater in northern India needs to be sustainable. But drought should not be induced as part of the solution.

Saumitra Mukherjee School of Environmental Sciences, Jawaharlal Nehru University, New Delhi 110067, India
e-mail: saumitramukherjee3@gmail.com

OPINION

Biodiversity

Costing the Earth

The value of biodiversity must be accounted for, says **Pavan Sukhdev**. It is time for governments to invest to secure the flow of nature's 'public goods'.

Clean air, fresh water, the flood protection provided by wetlands, the carbon-storage capacity of forests: these are examples of natural systems and processes that we largely take for granted. We consider them 'public goods': they are available to everyone; there is enough to go round; and one person's enjoyment of them does not impede another's. They are not traded in markets, not priced and they are mostly available for free.

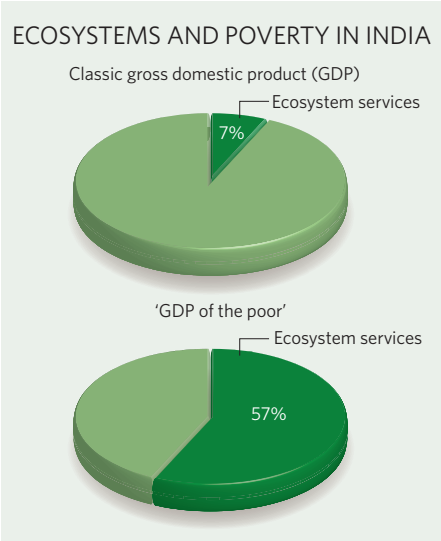
This attitude, held the world over by everyone from consumers to policy-makers, demonstrates a lack of understanding about the finiteness and fragility of ecosystem services. Their contribution to national gross domestic product (GDP) and to human well-being is barely recognized. The inevitable outcome is a situation described by ecologist Garrett Hardin more than 40 years ago as the 'tragedy of the commons', in which individuals who consume a shared resource according to their own self-interest are bound to destroy it (G. Hardin *Science* **162,** 1243–1248; 1968).

The tragedy of the commons is now greater than ever. Many natural resources are under increasing pressure from over-exploitation and changes of land use and yet still are available free of cost. The main problem is that ecosystems and biodiversity provide both private and public goods. For example, a logging company might pay for the right to harvest trees, but no one provides compensation for the loss of carbon storage that results from felling them.

Furthermore, many natural resources are 'open access' and not covered by property rights or effective national laws or international treaties, which leads to their constant depletion. For example, open access and a perverse system of subsidies have left two-thirds of fish stocks across the globe over-exploited, and have damaged coastal ecosystems. This threatens both the fisheries industry, which generates income of US$80 billion–$100 billion annually, and the livelihoods of 27 million people who depend on fisheries, most of whom are poor, small-scale fishermen. Additionally, more than a billion people, mainly in developing countries, rely on fish as their main or sole source of animal protein.

There is a pattern here: poor people are almost always hit hardest by the misuse of environmental resources as they depend on them most heavily. A recent attempt by the Green Indian States Trust (GIST), a non-governmental organization that promotes sustainable development, to develop a 'GDP of the poor' in India is a good illustration. The trust showed that although the value of forest services such as fresh water, soil nutrients and non-timber forest products was only around 7% of national GDP, it amounted to some 57% of the income of India's rural poor people (see graph).

There are many calls for changes to the current economic paradigm to solve this problem of declining public goods. One is through TEEB (the Economics of Ecosystems and Biodiversity), a major global study to draw attention to the tangible benefits of biodiversity, and to highlight the growing costs of biodiversity loss and ecosystem degradation (www.teebweb.org).



SOURCE: GIST GREEN ACCOUNTING FOR INDIAN STATES PROJECT, 2002-03 DATA

There are several realistic approaches on the table, including those developed by this year's joint-winner of the Nobel prize in economics, Elinor Ostrom. In her studies of resources such as forests, lakes, fish stocks and pastures, Ostrom found many cases in which communities have developed sophisticated mechanisms for the successful management of common property (E. Ostrom *et al. Science* **284,** 278–282; 1999).

Indeed, there is a wealth of initiatives around the world that work. In Costa Rica, for example, payments for environmental services have become virtually a countrywide strategy for forest and biodiversity conservation. They are funded largely from transportation taxes, directly benefit farmers who resolve to retain forest patches on their lands and indirectly benefit many others who enjoy the continuing 'public' benefits from these forests. Companies are also increasingly seeing value in biodiversity preservation and recognizing the interconnectivity with long-term business durability. Insurance firms and shipping companies have financed the reforestation of the Panama Canal region to restore freshwater flow to its locks and prevent the rise of shipping premiums caused by the risk of canal closures.

It is now up to governments to provide fiscal or other incentives to encourage the participation of stakeholders as responsible stewards rather than short-term optimizers. This can be done by reforming the way property and access rights are assigned, and through better targeting of taxes and subsidies. Globally there are many unused opportunities.

One game-changing mechanism being discussed at the UN climate conference in Copenhagen later this year is REDD+, a proposed scheme to reward reductions in deforestation and degradation, as well as rewarding more afforestation or reforestation, and effective conservation. This links the maintenance of ecosystems with the climate challenge by setting up a global scheme to reduce emissions from forest losses (which are close to a fifth of all emissions) and to encourage more and better carbon capture through forestry. Many countries are already developing capacity to implement REDD+. The upside of such preparation is significant — it could set up a framework that might be used to reward other 'public goods' such as freshwater capture and storage, species conservation and community livelihoods.

There are signs that countries are willing to put new economic models in place to stop the misuse of their environmental resources and to limit degradation. Although many uncertainties remain, good ideas for change are close at hand. We just have to lean forwards and pick them up. ■

Pavan Sukhdev is study leader of the Economics of Ecosystems and Biodiversity (TEEB) project, United Nations Campus, Hermann-Ehlers-Strasse 10, 53113 Bonn, Germany.
e-mail: teeb@unep-teeb.org

See Editorial, page 251, News Feature, page 270, and biodiversity special at www.nature.com/darwin. To hear an interview with the author see go.nature.com/oLMXVt

Biodiversity

A force to fight global warming

Natural ecosystems and biodiversity must be made a bulwark against climate change, not a casualty of it, argue **Will R. Turner**, **Michael Oppenheimer** and **David S. Wilcove**.

In the tortured history of climate-change negotiations, enlightened thinking has translated into positive action all too rarely. But governments have recently seen the light on a crucial issue: they have recognized the vital role that intact natural ecosystems have in limiting the build-up of atmospheric greenhouse gases.

When delegates convene in Copenhagen next month to strengthen the UN Framework Convention on Climate Change (UNFCCC), an initiative to preserve the world's forests to store and sequester carbon will take centre stage. Reducing emissions from deforestation and forest degradation (REDD) should give developing countries the opportunity to benefit financially by preserving their forests, either through direct payments or by allowing them to market the carbon stored in uncut trees. Its backers hope that with sufficient funding REDD could substantially slow rates of deforestation, especially in the tropics.

REDD is just one of many possible ways to exploit the potential of natural ecosystems to slow climate change and lessen its effects on people. Natural habitats are a hugely valuable tool in the fight against global warming. Use them wisely and they could save many lives and vast sums of money in the decades to come. Abuse them, and much of Earth's biodiversity could be lost, along with the fight against climate change. Urgent action is needed to understand how best to exploit this promise and develop mechanisms that can be woven into the practices of governments, corporations, communities and institutions worldwide.

> "Climate change is seen as one problem for nature and another for people. This must stop."

To achieve such an integrated approach means fighting a host of powerful short-term political and economic interests. The carbon markets created by REDD might invite corruption, as many critics suggest. Yet the rapid progress that has already been achieved in anticipation of REDD — including new financial mechanisms to ensure verified and lasting emissions reductions, and innovative remote sensing and mapping tools to support them — suggests that these challenges are surmountable[1].

There are two good reasons for focusing on natural ecosystems for tackling the threats of climate change. First, forests, peatlands, oceans and other ecosystems control carbon and other global biogeochemical cycles. The oceans alone sequester about 2 gigatonnes of carbon a year. Reducing deforestation and forest degradation rates would slash global emissions by up to 1 gigatonne of carbon a year, more than the emissions of all passenger cars combined. Restoring the world's marginal and degraded lands to natural habitats could sequester an additional 0.65 gigatonnes annually.

The second reason has to do with practicality: the maintenance and restoration of natural habitats are among the cheapest, safest and easiest solutions at our disposal in the effort to reduce greenhouse-gas emissions and promote adaptation to unavoidable changes (see graphic). The basic materials already exist — so there is no need for technological development. Indeed, ecosystem restoration (for example, replanting forest on previously cleared land) may remain for several decades the only realistic large-scale mechanism for removing carbon dioxide already in the atmosphere[2].

Natural protection

Environmental carbon storage is worth trillions of dollars to the world's economies, yet it is only one of nature's services. Natural ecosystems will save lives and sustain livelihoods in myriad ways as Earth's climate changes[3]. For example, healthy mangroves, reefs and wetlands can protect people and property in coastal and inland communities even as climate change threatens to increase tropical cyclone activity. A cyclone in Orissa, India, in 1999 would probably have killed three times as many coastal residents if mangrove forests had not buffered their villages[4]. Even at current storm levels, coastal wetlands in the United States provide protection against hurricanes worth an estimated US$23.2 billion a year[5].

Natural ecosystems do many climate-related jobs. Mangroves, for example, store carbon, buffer against storm impacts, support fisheries and harbour diverse species. Ecosystems also support livelihoods by providing alternative sources of income and food, especially useful if climate change disrupts current sources. Such diversification is helpful for everyone, particularly for the most vulnerable countries and communities — those with the least capacity to cope with climate change.

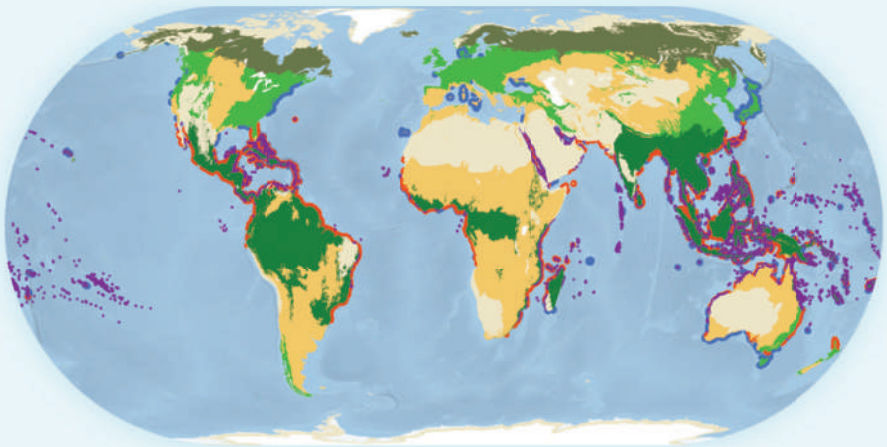
As important as these services are, what remains to be discovered may be more valuable still. Three decades ago, few imagined that the carbon stored in natural systems would become crucial for combating climate change. Today, enzymes from the gut of a marine crustacean (*Limnoria quadripunctata*), a type of gribble, show promise in breaking down agricultural waste products for biofuels, potentially reducing greenhouse-gas emissions without competing for agricultural land or threatening natural habitats[6]. If a promising biotechnology can emerge from a common woodlouse-like creature that lives on the underside of a busy British pier, what untapped potential — the 'option value' of biodiversity — might lie in the world's wildernesses? One area where this untapped innovation could prove particularly valuable is agriculture. When changes in precipitation and temperature start to test the physiological limits of current crops, farmers could benefit from wild relatives and novel cultivars better suited to the new conditions.

The danger is that we will overlook these benefits in natural systems or, worse, lose them. Vast areas of wilderness and undeveloped land are already falling to human abuse, either directly via habitat destruction or indirectly through the effects of climate change. One-fifth of all vertebrates are now threatened with extinction[7], and habitat destruction is estimated to cost $2 trillion–5 trillion annually in lost ecosystem services such as the provision of water and carbon storage, vastly more than the cost of safeguarding those services.

Halting this decline requires identifying and securing key intact ecosystems and the climate services they provide, restoring lost or degraded ones, and limiting future losses, all in partnership with the communities that need those services most. At present, climate change is seen as one problem for nature and another for people. This must stop. If human adaptation to climate change compromises biodiversity, then the loss of forests and other natural ecosystems will accelerate climate change, increasing the need for adaptation even as the planet's capacity to accommodate it diminishes. An integrated approach makes the circle virtuous: by conserving biodiversity, we decelerate climate change while increasing the adaptive capacity of people and ecosystems alike.

THE BENEFITS OF BIODIVERSITY

The maintenance and restoration of natural habitats are among the cheapest, safest and easiest solutions that could aid the effort to reduce greenhouse-gas emissions and promote adaptation to unavoidable climate change.



Biome	Percentage of habitat lost	Percentage of species threatened	Examples of benefits for climate-change adaptation/mitigation
Tropical forests	30	22	Avoided carbon emissions, water regulation/purification, reduced erosion and flooding, biodiversity 'option value'
Temperate forests	48	15	
Boreal forests	2	5	
Grasslands, shrublands and Mediterranean habitats	57	17	Water regulation/purification, crop wild relatives
Deserts and tundra	21	12	Avoided carbon emissions from soil
Seagrasses	29	16	Carbon absorption and avoided emissions, limiting coastal erosion, fisheries stability, reducing cyclone impacts, biodiversity 'option value'
Mangroves	37	11	
Coral reefs	20	27	

Sources: WWF; UNEP-WCMC; Waycott, M. *et al. Proc. Natl Acad. Sci. USA* **106,** 12377–12381 (2009); Millenium Ecosystem Assessment; IUCN Red List

There is a real possibility that Copenhagen will create a mechanism for REDD but not a means to pay for it. So the parties to the UNFCCC must initiate financial incentives immediately, engaging public and private sources of funding so that REDD can be rolled out on a global scale. Action is also needed to help governments to monitor natural and modified ecosystems for true net emissions — including those that arise through the displacement of food crops by biofuels and other land-use changes. Policies that provide benefits to, and respect the rights of, local communities are crucial for sustaining and enhancing the ability of natural ecosystems to mitigate climate change.

We also need to try to find ways to value and market the other climate services that natural habitats provide — acknowledging that such services do not exist everywhere — and to weave these benefits into the fabric of our economies. For example, residents of Quito, Ecuador, pay via their utility bills to protect upstream habitats that provide much of their fresh water. Yet, in most cases, communities and corporations are either unaware of, or ignore, the degree to which aspects of climate security, such as their water supply, depend on natural ecosystems.

Climate reserves

For centuries, people have safeguarded natural habitats as public parks and privately owned reserves for nature conservation, sustainable resource production and other goals. They must now be harnessed for the additional goal of combating climate change. National governments, including Costa Rica and the United States, have already begun to acknowledge the importance of natural ecosystems for adaptation in their submissions to the UNFCCC. They need to ensure that any agreement emerging from Copenhagen contains substantive measures to promote the conservation of these ecosystems. Parties to the UNFCCC must also develop key principles for managing and restoring climate services that can be incorporated into international treaties, such as the Convention on Biological Diversity's inland waters biodiversity programme, now under development, as well as environmental assessments by development banks. Knowledge and resources for harnessing climate services should be shared internationally, with developing countries being supported by developed countries.

The future of economies and livelihoods across the planet depends on integrating biodiversity conservation into climate-change planning. If REDD is allowed to fail and degraded lands are also not restored, it is likely to be very difficult to avoid dangerous temperature increases. If coastal and wetland ecosystems are not preserved and restored, tropical storms will become more deadly and more economically damaging. If the diversity of life in the world's wildlands and waters disappears, so do eons of natural innovation that could yield breakthroughs. Working with natural systems rather than against them would unleash a powerful, essential force for halting climate change and reducing its impacts. ■

Will R. Turner is director of global priorities at the Center for Applied Biodiversity Science, Conservation International, 2011 Crystal Drive, Suite 500 Arlington, Virginia 22202, USA. **Michael Oppenheimer** is Albert G. Milbank professor of geosciences and international affairs and **David S. Wilcove** is professor of ecology and evolutionary biology and public affairs at Princeton University, Princeton, New Jersey 08544, USA.
e-mail: w.turner@conservation.org

1. Transparency International *Global Corruption Report 2009* (Cambridge Univ. Press, 2009).
2. Hansen, J. *et al. Open Atmos. Sci. J.* **2,** 217–231 (2008).
3. Locke, H. & Mackey, B. *Int. J. Wilderness* **15,** 7–13 (2009).
4. Das, S. & Vincent, J. R. *Proc. Natl Acad. Sci. USA* **106,** 7357–7360 (2009).
5. Costanza, R. *et al. Ambio* **37,** 241–248 (2008).
6. *Sailors' historic scourge may hold the key to bioenergy future* (Press release, Univ. York, 2009); available at www.york.ac.uk/news-and-events/news/2009/gribbles-bioenergy
7. IUCN *2008 IUCN Red List of Threatened Species* (2008); available at www.iucnredlist.org.

The authors declare competing financial interests: details accompany the article online at go.nature.com/9alss1.

See Editorial, page 251, and News Feature, page 266. For the whole biodiversity special, see www.nature.com/darwin.

Biodiversity

Let the locals lead

To save biodiversity, on-the-ground agencies need to set the conservation research agenda, not distant academics and non-governmental organizations, argue **Robert J. Smith** and colleagues.

To stem biodiversity loss, it is essential to identify priority areas for conservation and take effective action within them. However, much of the research on this topic is only peripherally relevant to these tasks, and contributes little to local conservation efforts. Researchers focus on their pet interests and on making an impact in the scientific literature, taking little notice of the institutions and organizations that actually develop and implement conservation plans. Meanwhile, international non-governmental organizations (NGOs) influence academics' priorities without the latter always appreciating the constraints that shape NGO agendas.

A good example of how global NGOs, scientific fashion and academic journals combine to marginalize relevant conservation science is the noisy and largely fruitless debate that followed the publication of a biodiversity hotspots map[1] in 2000 by Conservation International, a leading NGO. The map was marketed as a tool for identifying where conservation investment would have the biggest impact, but this involved playing down both how little was actually known about species distributions and that accurate global data sets on the costs of implementation were not available.

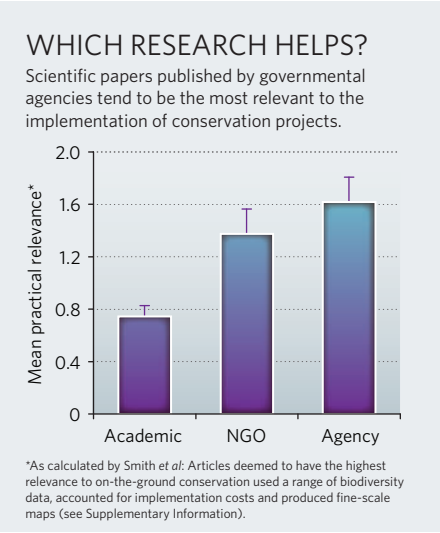
These limitations did not stop the map doing its main job, which was to raise funds and show broadly where Conservation International should target its efforts. In fact, the initiative has been extremely successful and helped to raise an estimated US$750 million for conservation within hot spots. But the hype led many academics to treat priority area setting as simply a question of working out what lives where. This led to many studies that took no account of how plans are implemented.

Different priorities

International NGOs have played a vital part in identifying important conservation areas, especially in countries where local government agencies are underfunded or ineffective. Yet because each NGO answers to its members and donors, its priorities can never exactly match those of local conservationists (see graphic). The constant pressure on NGOs to fund-raise and market their work exacerbates this disconnect. When local institutions are weak, this can cause three problems.

First, the need to create a sense of urgency among donors leads to short-term funding and 'quick and dirty' projects, which rarely gain local long-term support[2]. Second, NGOs tend to advocate their institutional methodology, rather than allowing local agencies to develop approaches that best match their needs. Third, NGO researchers find it easier to produce articles on broad-scale issues for high-impact journals, which helps to build scientific support for new campaigns[3], than to write papers about research on local issues.

Saving globally important biodiversity requires a radical rethink by conservation scientists. Researchers must allow government conservation agencies and other local groups to set the broad agenda for research and decide how to implement results. International NGOs should have only a supporting role: conservation plans are more legitimate and politically acceptable when set locally[4], where they can also be better coordinated with other sectors such as land-use planning, agriculture, water and climate change.



The advantages of this approach can be seen in the South African province of KwaZulu-Natal (KZN). Ezemvelo KZN Wildlife, the agency charged with conserving biodiversity and managing the province's network of protected areas, works with researchers to identify areas for conserving nationally and internationally important species, habitats and ecosystems[5], helping match donor and NGO priorities with those of the province. Crucially, this prioritization forms part of a broader planning scheme, which is continually updated and used to monitor the effectiveness of protected areas and assess development applications. This system has guided the South African government's National Protected Area Expansion Strategy, the WWF-Netherlands Black Rhino Range Expansion Project and the proposed Critical Ecosystem Partnership Fund programme in the Maputaland-Pondoland-Albany hot spot.

Local approach

South Africa is a world leader in priority area setting and implementation[6]. But agencies in many other developing countries lack the means and influence to implement change. How should governments and researchers overcome such shortcomings? One way is to develop 'social-learning institutions', which bring together local and international conservationists and researchers. Government staff are often poorly trained, funded and motivated. Working with outside experts gives them access to new skills and contacts, enabling them to develop their own conservation agendas. This is essential at a time when too many conservation programmes in biodiversity-rich developing countries are driven by foreigners, an approach that causes local resentment and makes conservation seem a marginal issue.

An example of an effective social-learning institution is the Thicket Forum in South Africa's Eastern Cape province. The Forum brings together representatives from government, NGOs, consultancies, farmers and other landowners with ecologists and social scientists to exchange knowledge and identify priorities for research and training[7]. Academics have a crucial role, but the entire membership sets the agenda, which helps in balancing views and targeting activities. This has been particularly useful for guiding research, and for training the next generation of conservation professionals. Such an approach is not unique to South Africa or the conservation sector and could be widely adopted elsewhere.

To help ensure that local organizations are equipped to make decisions about conservation research, foreign donors must collaborate directly with them about their specific requirements. Donors should also fund local

M. POWELL

The Working for Woodlands project assesses the survival of cuttings planted to restore degraded thicket in the Baviaanskloof World Heritage Site in South Africa.

groups directly, to enable them to finance the establishment of social-learning institutions and the research priorities that those groups identify, and to train agency staff and local experts. Currently, donor money tends to flow through international NGOs, so in many countries weak local agencies stay that way. An independent oversight body is needed to coordinate such direct funding, to ensure that responsible local agencies receive consistent funding from donors.

Once such mechanisms are in place, what kind of research should social-learning institutions and local conservation agencies do? There is no shortage of topics: investigating what makes people support or block conservation projects; the social and economic implications of different methods for prioritizing conservation areas; the social, economic and biodiversity benefits of different management approaches; the effectiveness of these conservation projects; approaches for building support for conservation; and many more. Research on the geographical distribution of biodiversity and how this will change with climate change — one of the hottest topics in academic conservation science — is also needed in some places, but only if produced in a way that helps local decision-making.

Finally, international organizations and journals can help to refocus conservation research on what local agencies and communities need. For example, the World Commission on Protected Areas, which represents the interests of governments and NGOs, could catalogue the approaches used around the world and highlight the most successful, so local agencies can see what has worked and what hasn't. The United Nations Environment Programme World Conservation Monitoring Centre could make national maps of priority areas internationally available. This would be social learning on a grand scale. In addition, research journals should assess whether articles make naive assumptions about implementation, and recognize the value of locally important case studies, rather than assuming that global analyses are always the most powerful.

Such profound changes would need to be made carefully. Directly funding weak agencies has often failed in the past and we need to learn from those mistakes. The new approach should be piloted in countries where success is most likely and then adapted to more challenging locations. Some conservation agencies will initially need a lot of help from international NGOs and researchers, but this should not be a problem as long as the financial and academic incentives are in place.

> **"If academics really want to achieve results, they should join or set up social-learning institutions."**

The conservation-science community should recognize those with the highest academic or media profile can no longer set the research agenda. Moreover, academics need to understand that if they work in isolation from local conservation agencies, those who might usefully apply their research will probably ignore it. If academics really want to change the conservation agenda or achieve results on the ground, they should join or set up social-learning institutions as part of a planning process. This will take more time than simply firing off another paper, but it will also lead to more interesting, novel and important research. ■

Robert J. Smith, Diogo Veríssimo, Nigel Leader-Williams are at the Durrell Institute of Conservation and Ecology, University of Kent, Canterbury, Kent CT2 7NR, UK.
Richard M. Cowling is in the Department of Botany, Nelson Mandela Metropolitan University, Port Elizabeth 6031, South Africa.
Andrew T. Knight is in the Department of Conservation Ecology and Entomology, Stellenbosch University, Matieland 7602, South Africa.

1. Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. & Kent, J. *Nature* **403,** 853–858 (2000).
2. Knight, A. T. *et al. Bioscience* **57,** 256–261 (2007).
3. Smith, R. J., Veríssimo, D. & MacMillan, D. C. in *Trade-offs in Conservation: Deciding What to Save* (eds Leader-Williams, N., Adams, W. M. & Smith, R. J.) (Wiley-Blackwell, in the press).
4. Rodriguez, J. P. *et al. Science* **317,** 755–756 (2007).
5. Goodman, P. S. *Bioscience* **53,** 843–850 (2003).
6. Balmford, A. *Trends Ecol. Evol.* **18,** 435–438 (2003).
7. Knight, A. T. & Cowling, R. M. *S. Afr. J. Sci.* **102,** 406–408 (2006).

Supplementary information accompanies this article online at go.nature.com/hrKvfg.

See Editorial, page 251, and News Feature, page 266. For the whole biodiversity special see www.nature.com/darwin.

Biodiversity

A call to the custodians of deep time

Palaeontologists must model the causes of biodiversity rather than simply cataloguing fossils, says **Douglas Erwin**, as they curate the only record of ecosystems undamaged by humans.

Humans have been destroying much of Earth's life for the past 10,000 years. It might be comforting to think that we have a reasonable grasp of the factors lying behind the planet's number of species and their abundance, should we ever choose to reverse this course. But to a large extent, we don't.

Today, ecologists can look at the biodiversity of tropical forests and compare that with polar regions, listing the reasons for the differences they find. But in the fossil record the picture of diversity is more confused. Palaeontologists cannot even agree whether diversity was higher 10,000 years ago than it has ever been, or whether it plateaued hundreds of millions of years ago. The confusion persists because palaeontologists continue to simply look for patterns in the fossil record, which is incomplete and subject to interpretation. Instead, we should be seeking to study the basic processes that underlie diversity, and building models to test those theories. Only by understanding what governs diversity can we settle old arguments about how it has changed over time.

The debate over which scenario is correct has an impact on fundamental questions about how evolution builds biodiversity and the importance of evolutionary innovations. Has the ecological space occupied by Earth's biota expanded over time, or has it simply become more crowded? Does intra-species competition keep diversity in check, or does the creation of one species construct new environments and opportunities for others, allowing diversity to bloom?

Today, palaeontologists are the custodians of the only record of ecosystems undamaged by human activities. In 2006, palaeoecologists found that modern, human-modified Caribbean reef ecosystems are completely unlike any reef community of the past 220,000 years[1]: studying modern Caribbean reef ecosystems leaves us blind to how they originally operated. The same is true, I am sure, of other ecosystems. Although the fossil record is far from perfect, it is perhaps the best resource for understanding the processes involved in the development of biodiversity, and the only record that we can use to evaluate the long-term significance of competition and evolutionary innovation. To do so, those interested in the history of life must move beyond collecting fossils to creating models.

Life on Earth is some 3.5 billion years old. Before oxygen levels rose in the ocean and atmosphere some 600 million years ago, global diversity was probably limited. The macro-fossil record begins in earnest with an explosion of animal life some 575 million years ago — just before the start of what is called the Phanerozoic eon around 540 million years ago. This culminated in the 'Cambrian explosion' (535–520 million years ago). Since then there have been several periods of apparent rapid increases in diversity, as well as stunning mass extinctions that eliminated many evolutionary lineages. All of the non-avian dinosaurs, for example, were killed at the end of the Cretaceous period 65 million years ago.

The fossil record is dramatically incomplete, at all scales. In the metaphor of Charles Lyell and Charles Darwin, it is as if we have only a few chapters of a book, and but a few pages of each chapter with only some fragmentary sentences on each page. I would say the record is a bit better than that, but it remains true that in any one place there is more time missing than there is preserved.

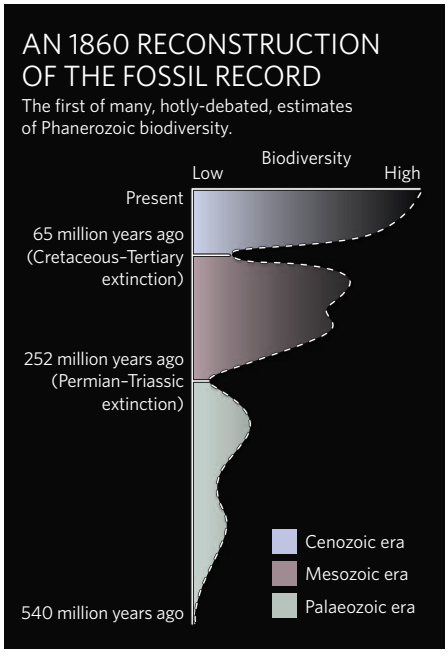
> "Only by understanding what governs diversity can we settle arguments about how it has changed."

The vagaries of the fossil record have led to differences in interpretation. Since Oxford palaeontologist John Phillips first chronicled the fossil record in 1860 (see graphic, below), most have assumed that diversity has grown over time, particularly over the past 100 million years. But in 1972, palaeontologist David Raup of the University of Chicago argued that marine diversity might well have reached a plateau over 400 million years ago; the apparent increase since then, he argued, was an illusion created by increasingly better preservation of fossils in younger rocks.

A 1981 paper in *Nature* by Raup and many of the other protagonists in this debate led to an uneasy consensus in favour of strong increases in diversity in the early Palaeozoic (542–252 million years ago), and again over the past 100 million years[2]. For several decades Raup's concern bubbled away in the background as palaeontologists compiled steadily more detailed databases of the durations of fossil families and genera, based on the oldest and youngest known samples of each. The most comprehensive of these was constructed by the late Jack Sepkoski, also of the University of Chicago, and includes more than 37,000 genera.

Sepkoski concluded that there was a roughly linear, several-fold increase in diversity through the Phanerozoic, which he believed would ultimately level out[3]. Others analysing the same data argued that global diversity had expanded exponentially, particularly over the past 100 million years or so. In July 2008, results from the Paleobiology Database project (http://paleodb.org), which aims to collect a record of all fossil finds, built up location by location, and includes 3.5 million records, suggested that recent marine biodiversity is only 1.5–1.8 times the average of the Palaeozoic[4] — much less than some expectations.

In the absence of any expectations about what sort of diversity pattern we should expect to find, it is difficult, if not impossible, to tell which reconstruction of Phanaerozoic diversity is correct.



J. PHILLIPS *LIFE ON THE EARTH: ITS ORIGIN AND SUCCESSION* (MACMILLAN, 1860)

Ecological spaces

Palaeontologists often say that a burst of diversity in the fossil record simply 'filled in ecological space', as if each new species simply took up residence in a square of a pre-existing chessboard. This would match Darwin's notions of intra-species competition being the main driver behind how life fits into ecological niches, and would imply that diversity should grow only slightly over time.

I think a much better analogy is of building the chessboard itself. Although some of these ecological spaces may exist independently of any species that occupies them, many more are defined by species and their mutual interactions. I would argue that the chessboard grew during the Cambrian diversification, for example, increasing its overall ability to support life. Thus we should expect significant increases in biodiversity over time, particularly when adaptive breakthroughs open up new opportunities or construct new habitats such as reefs.

How does such a chessboard grow? As a first step in answering this question, first-order, process models of global biodiversity need to be developed that are informed by our understanding of past climates, continental configurations, and geochemical cycles. For example, diversity is much higher near the equator than near the poles. So we might predict there would also have been greater diversity at times in the past when the area of tropical marine shelves was greater, or during global warm periods. However, warm periods might also lead to nutrient-trapping in continental basins and the spread of oxygen-free waters, thus restricting diversity. The number and position of continents, each carrying its own ark of species, is another important variable. Basic models should be built in which the input of such variables can reliably predict biodiversity at various places and times.

What spurred brachiopods to such diversity?

SMITHSONIAN INSTITUTION/D. ERWIN

Additions will then be needed to address ecological interactions. Ecologists have begun work on models of diversity, some of which have spurred an ongoing debate about the importance of niches in biodiversity. Ecologists have also examined how organisms such as corals construct habitats for other organisms, thus boosting diversity. But ecologists have only begun understanding these processes over the past decade, and palaeontologists have done very little to explore their application to deep time. My colleagues and I, for example, are studying the role of ecosystem engineering in the Ediacaran–Cambrian diversification of early animals (579–510 million years ago). My suspicion is that some evolutionary innovations have disproportionate effects on diversity — the advent of burrowing behaviours, for example, might have changed marine sediment chemistry and microbial productivity in ways that spurred further innovation. But for now we have only qualitative arguments.

Models built to examine these ideas will need to be tested empirically, so we need to ensure that data are being collected in an appropriate way for future use. If, for example, a model is built to look at the effects of filtering sponges and clams on water quality and subsequent diversity, then simply knowing that those animals lived at a given place and time will be insufficient: the number of sponges and clams will also need to be determined. But this information isn't always captured. One of the world's largest palaeobiology collections, for example, contains more than a million fossils gathered by G. Arthur Cooper and Richard E. Grant in west Texas from the 1930s to the 1980s, but doesn't contain sufficient information to determine the relative numbers of fossils in the various stratigraphic layers. This is a huge loss to the field. Today, palaeontologists are more likely to count the numbers of fossils they find in each rock layer and use the same approach from one location to another. But this is time-consuming, and not always done. This needs to become standard.

Any understanding of dynamics also requires a firm knowledge of when things happened in the fossil record, and for how long. Very-high-resolution radiometric dating (accurate and precise to 0.1% or better) is an appropriate tool (see EARTH-TIME.org).

Model culture

Such model building would come naturally to physicists. When they designed and built an instrument to look for the Higgs Boson, for example, they did not plan to simply turn it on and sift through all the data; they developed models to help them predict what they might see if certain theories about the Higgs were true, and they will match their collected data to their predictions. Palaeontologists do not tend to think like this.

Palaeontologists share a common interest in the dynamics of growth with economists. Although the global recession has dealt a blow to economic modelling, basic lessons can still be learned from the beginnings of the field. Early models of economic growth were incredibly simple, focusing on the numbers of workers, for example, without any consideration of how those workers might interact. When combined with empirical data, the models provided great insights into economic growth (and won their developer a Nobel prize). In these early models, sustained economic growth was achieved only by adding an arbitrary variable for technological innovation; economists eventually unpacked the processes of economic growth by opening this 'black box' of technological innovation and deconstructing the processes that drive it. The lesson to learn is that relatively simple equations, if properly constructed, may explain much of an observed pattern, allowing researchers to focus on the unexplained portion.

Palaeontologists need to adopt a similar approach. Building simple, 'toy' models of biodiversity will be relatively easy. Developing more sophisticated approaches will be a job not just for palaeontologists, it will also require the input of geologists and, more importantly, physicists and theoretical biologists.

My palaeontological colleagues may object that diversity is far too complicated, but that is of course the point of good process-oriented models, for they extract from complexity the critical variables that are believed to control the dynamics. Through iteration with empirical testing the models will improve, and provide an important spur to future research.

To get the full picture of biodiversity, we must explore the fossil record. But we must do so deeply — burrowing beneath patterns in search of the processes that control diversity, and building models to test these ideas. ■

Douglas Erwin is in the Department of Paleobiology, National Museum of Natural History, Washington DC, USA and at the Santa Fe Institute, Santa Fe, New Mexico, USA.
E-mail: erwind@si.edu

1. Pandolfi, J. M. & Jackson, J. B. C. *Ecol. Lett.* **9,** 818–826 (2006).
2. Sepkoski, J. J. Jr, Bambach, R. K., Raup, D. M. & Valentine, J. W. *Nature* **283,** 435–437 (1981).
3. Sepkoski, J. J. Jr *Paleobiology* **10,** 246–267 (1984).
4. Alroy, J. *et al. Science* **321,** 97–100 (2008).

See Editorial, page 251, biodiversity special at www.nature.com/darwin and palaeontology podcast at go.nature.com/UuNpgJ

Global Darwin: Multicultural mergers

Latin Americans first saw evolution as a reason to 'whiten' their societies, then as a reason to take pride in their mixed lineage, says **Jürgen Buchenau** in the last of four pieces on Darwin's global influence.

Darwin200

On 28 February 1832, the HMS *Beagle* arrived at the port of Bahia, Brazil, its second stop on a five-year exploration of the globe. The impressions that the continent's peculiar animal and plant life made on the 22-year-old Charles Darwin have been well documented. Less well known is the effect Darwin had on the people of Latin America.

In the late 1800s, Latin American intellectuals, many of whom were politicians, used Darwin's ideas to promote mass immigration from Europe to 'whiten' and so 'evolve' their people. Some 50 years later, Latin American thinkers abandoned this emphasis on European superiority and instead supported the racial mixing, education and unification of the region's existing populations. That the social implications of evolution were interpreted so differently in such a short period of time is testament to the extraordinary ability of people to bend Darwin's ideas to fit ever-changing intellectual and political contexts.

Darwin's life and work coincided with the beginnings of modernization of the larger Latin American nations after a long period of chaos. By 1859, when *On the Origin of Species* was published, countries such as Argentina, Brazil and Mexico had suffered decades of economic stagnation and political instability after achieving independence from European countries in the 1820s. Their largely Catholic societies were made up of African American and Amerindian majorities dominated by a small white elite.

Throughout the continent, this elite tended to divide into liberals and conservatives, with each group having different takes on how to improve society. The conservatives, wary of Protestant nations such as the United States, favoured the development of internal economic markets. The liberals, many of whom had studied overseas, believed that foreign investment was key to building their countries' infrastructure. In the second half of the nineteenth century, when industrial production skyrocketed in western Europe and North America, liberals gained power throughout Latin America, and directed the destinies of most nations until the Great Depression of the 1930s.

As well as developing export economies to feed Western demand for raw materials, liberal politicians sought to evolve their societies based on their own version of social Darwinism. They soaked up the latest ideas from Europe, and read the works of philosophers such as Herbert Spencer and Francis Galton, Darwin's cousin and the inventor of eugenics. Most Latin Americans thought that society, like nature, evolved from primitive to complex structures, and saw the industrial societies of western Europe as being more culturally sophisticated than their own. They maintained that Latin American societies could evolve towards the supposedly superior European and US models.

Modernizers held various views on how best to achieve this progressive change. Some embraced the 'hard inheritance' theory of German priest and biologist Gregor Mendel, and argued that 'whitening' their nations' stock through interbreeding was the only path to societal improvement. Others followed the 'soft inheritance' notion of French naturalist Jean-Baptiste Lamarck and countered that people's inheritable traits could be changed simply by altering their environment, including their education, diet and living conditions.

Initially, the Mendelians prevailed. Latin American governments attempted to recruit prospective European immigrants. The surge of Europeans that had entered the United States had shown the draw of available farmland in attracting foreigners. With scant funds, governments sponsored colonizing companies to send recruiters to Europe to lure farmers to underpopulated rural areas in Latin America.

> "Most Latin Americans thought that society, like nature, evolved from primitive to complex structures."

Those countries with ample farmland and a climate similar to that of western Europe succeeded in pulling people across. Between 1870 and 1930, more than 11 million Britons, Germans, Irish, Italians, Portuguese and Spaniards settled in Argentina, southern Brazil, Chile and Uruguay. By 1900, people of European origin dominated society in Argentina and Uruguay. Other nations with large indigenous populations and little available farmland — such as Mexico and Peru — saw much less immigration from Europe.

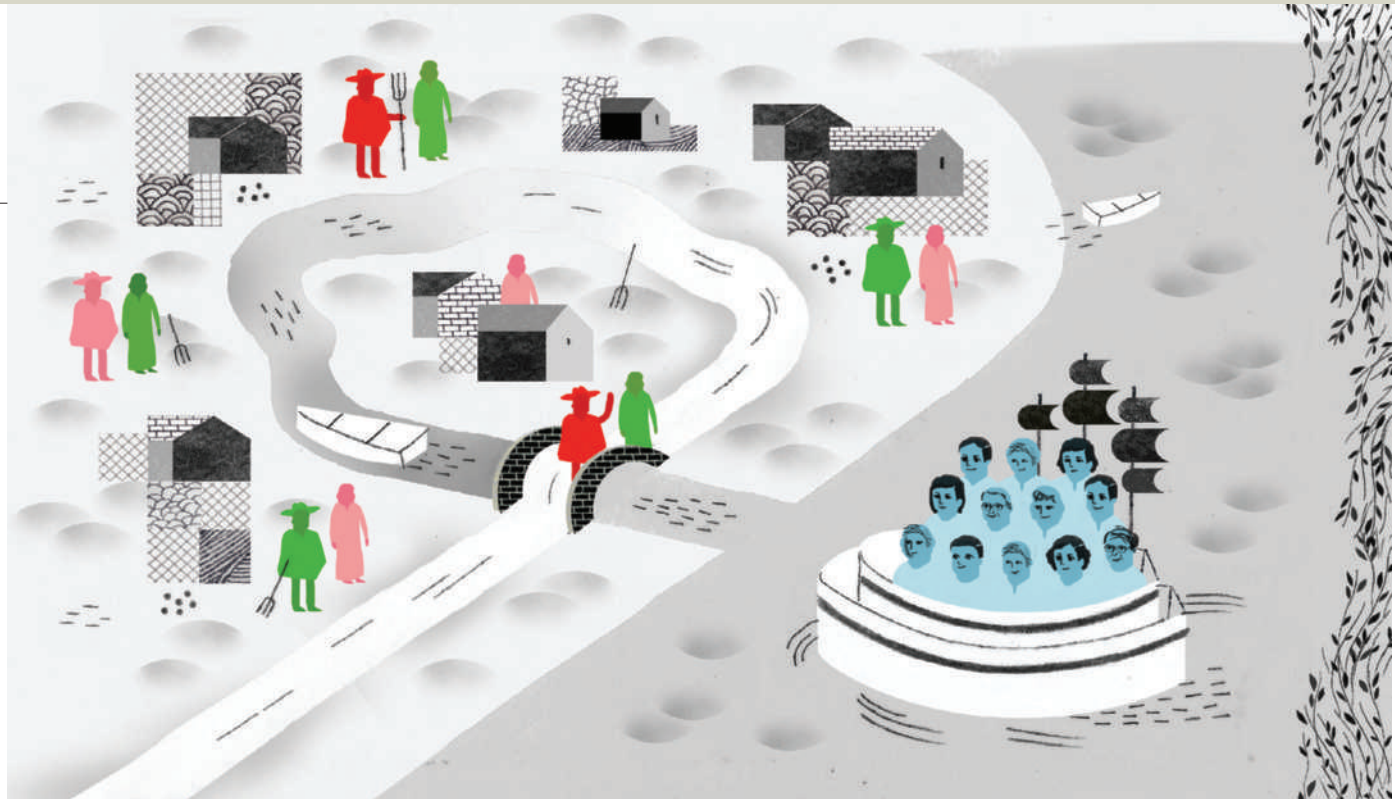
European ideas and values spread across Latin America at the expense of Amerindian and African American ones, with the establishment of European-style cities and institutions. For example, the government of dictator Porfirio Díaz remodelled parts of Mexico City, importing Italian marble to replace local building materials and following European trends in architecture and street design. They pushed the poor majority out of sight when foreign investors came to visit the capital.

No European superman

Attempts to westernize Latin Americans were short-lived: the tumultuous first half of the twentieth century destroyed the liberal modernizers' belief in European superiority — and with it, the idea of historical evolution through white immigration.

The death toll of the First World War demonstrated that Europeans had not evolved into superior human beings. A decade later, the Great Depression swept away the export economies underlying modernization in Argentina at least as much as it did in Mexico and Peru, belying the notion that the whitening of the population would lead to permanent social progress. And amid the global economic crisis, the rise of totalitarian systems in Germany and the Soviet Union stripped away most of the remaining admiration Latin American intellectuals held for European models.

Instead of rejecting Darwin's ideas outright, a new group of intellectuals — the cultural nationalists — came up with a revised formulation of how their societies should evolve. Although Darwin wasn't specifically invoked in such theories, his body of thought was still influential; so much so that the cultural nationalists might today be described as having adopted their own brand of social Darwinism. In particular, Mexican and Brazilian thinkers began to see the unique ethnic mixtures that shaped their nations as assets rather than liabilities — as long as those of African, European and Amerindian descent could be fused into a single culture. They increasingly began to tout the blending of racial groups as a way to forge new and improved social systems — societies that would not suffer from Europe's many ills. Not all agreed: for much of the twentieth century, Argentines continued



G. LAM

to hope that their largely European-populated capital of Buenos Aires would someday emerge as the Paris of the New World.

Instead of encouraging immigration and with that, greater technological and societal complexity, Latin American cultural nationalists wanted to unify society through public education. Education had been available to no more than 15% of the population in most countries, and cultural nationalists knew that the expansion of literacy among the poor, non-European majority was crucial to instilling a sense of national pride and citizenship. Even more importantly, they knew that literacy campaigns would give them an opportunity to promote an official version of history — one that emphasized racial and ethnic blending as a source of national pride.

The cosmic race

José Vasconcelos, Mexico's first cabinet-level education secretary and a university-educated man of European descent, was one person who adopted this viewpoint. He argued that traditions of the past were something people should be proud of but that progress in the present required eroding cultural diversity by means of mass education. His *La Raza Cósmica* (*The Cosmic Race*) essay, published in 1925, presented Mexican history as an evolutionary process that led from Aztec and European beginnings to the mestizo, a mixture of European and Amerindian ancestry. Vasconcelos thought that the mestizo was a 'cosmic race' that was superior to its component parts. He highlighted the cultural achievements of the pre-Columbian civilizations, but argued that social progress would come from assimilating their descendants into the mestizo identity. Vasconcelos designed a rural education programme intended to bring literacy to those who could not read, and Spanish to the Amerindian minority that spoke more than 60 different languages and dialects. It took decades for the programme to succeed.

In the 1930s, Brazilian sociologist Gilberto Freyre similarly proposed that Brazil was a "racial democracy" whose people considered racial blending an advantage rather than a disadvantage. In an argument that became known as Lusotropicalism, Freyre maintained that the Portuguese colonists who brought African slaves to Brazil were uniquely suited to survive in the tropics and that the subsequent intermixing had created a harmonious society that contrasted positively with the racism persisting in the United States.

Freyre's ideas, most famously laid out in his 1933 *Casa-Grande e Senzala* (*The Masters and the Slaves*), contradicted the stark realities in Brazil where social status was largely predicated on the degree of African ancestry, and Brazilians of European descent held virtually all important political and economic positions. Like Vasconcelos, Freyre can be interpreted as borrowing from Darwin in arguing for the melding of races as a positive evolutionary step. Freyre also did so in part to play down social problems.

After Vasconcelos and Freyre, Darwin-inspired thought in Latin America became unfashionable, in large part due to the effects of the Great Depression and the Second World War. Just as the First World War had eroded the idea of European superiority, the Second World War dealt a serious blow to notions of human history as a progressive process. After the war, Latin Americans still found themselves lagging behind the industrialized world in terms of economic development. Influenced by socialism, many Latin American intellectuals and politicians discarded the idea of evolutionary, gradual progress, embracing instead social revolution as the solution to the region's problems. Rather than imitating the Europeans and the United States, socialists saw the industrialized North Atlantic societies as part of the problem. In 1959, the Cuban Revolution set up Latin America's first communist government, and in the 1970s, socialist governments were established in Chile and Nicaragua.

From roughly 1870 to about 1930, Darwin's ideas resonated in Latin American political thought. During this relatively short period, Latin America's intellectuals went from thinking that evolution by natural selection explained European geopolitical superiority to using evolutionary models to promote cultural nationalism and the unification of the multiracial societies in which they lived. Either way, they found Darwin's notion of evolution useful in developing their ideas. As these nations struggled to unify their mix of indigenous, European and African populations, they saw themselves — as did their counterparts in the United States, Canada and Australia — as societies under construction from scratch. Throughout, Latin American political thinkers shared an optimistic belief that these societies could and would 'evolve' in a positive direction — whatever that direction might be. ■

Jürgen Buchenau is chair of the department of history at the University of North Carolina at Charlotte, 9201 University City Boulevard, Charlotte, North Carolina 28223, USA, and is the author of *Mexican Mosaic: A Brief History of Mexico*.
e-mail: jbuchenau@uncc.edu

See go.nature.com/5bHVBD for further reading. For more on Darwin see www.nature.com/darwin, or to discuss all four pieces in the Global Darwin series see go.nature.com/Figu8x.

BOOKS & ARTS

Bonds forged on the high seas

Shared experiences on global voyages linked Darwin and his fellow naturalists, explains **Alistair Sponsel**.

Darwin's Armada: Four Voyages and the Battle for the Theory of Evolution
by Iain McCalman
W. W. Norton/Simon & Schuster: 2009.
432 pp. $29.95/£20

J. W. CARMICHAEL/NATIONAL MARITIME MUSEUM

Dangerous research: Hooker's vessel *HMS Erebus* and its sister ship *HMS Terror* in the Antarctic.

Iain McCalman takes an unconventional tack among historians commemorating Charles Darwin's anniversary year. His aim is to show that the triumph of Darwinism in scientific and public debate following the publication of *On the Origin of Species* in 1859 was the result of a collective effort by a handful of career scientists from relatively unprivileged backgrounds. The bond between Darwin and the most important of these lieutenants — Joseph Hooker, Thomas Henry Huxley and Alfred Russel Wallace — was founded on their similar experiences as, in Darwin's words, "co-circumwanderers" of the southern oceans.

A cultural historian rather than a specialist in the history of science, McCalman is aware that the decision to take a long voyage was a conventional scientific career move in the early- and mid-nineteenth century. Finding passage to a remote portion of the globe was one of the few options available to young men of limited means who wished to devote their lives to the study of nature. Whereas the well-connected Darwin was offered a place on the *Beagle* unsolicited, the more common avenue for aspiring naturalists was to train as a physician and then to scrap for an appointment as a medical officer to a naval expedition.

Darwin200

It was in this manner that the 22-year-old Hooker joined James Clark Ross's Antarctic expedition as the surgeon's mate in 1839. Seven years later an equally callow Huxley embarked for a survey of the Great Barrier Reef as assistant surgeon on Captain Owen Stanley's *Rattlesnake*. These two at least had steady jobs and official dispensations to collect and study the plants and animals they found during their voyages. Wallace had to pursue science as an independent prospector, travelling to little-studied parts of Brazil and the Malay archipelago in search of desirable specimens that he could sell to collectors.

The years of sea-sickness, danger and tropical disease served all four men as an investment. As the unproven Hooker wrote when he was lobbying Captain Ross for a place on the *Erebus*, just three years after the *Beagle* had returned Darwin to Britain, "what was Mr. D[arwin] before he went out? He, I daresay, knew his subject better than I do now, but did the world know him? The voyage with FitzRoy was the making of him."

This book's claim to novelty, and its main virtue, lies in McCalman's decision to juxtapose the periods in each of the protagonists' lives when they were travelling. Individually, Darwin, Hooker, Huxley and Wallace are already the best-studied anglophone naturalists of the nineteenth century. McCalman breaks no new ground in his research, relying on their published travel narratives and memoirs and on recent biographies, particularly those by Janet Browne, Jim Endersby, Adrian Desmond, James Moore and Ross Slotten. But by holding events in England at arm's length, McCalman forces us to notice the similarities of the men's intellectual and emotional experiences as well as of their physical privations. He thus sheds light on the depth of their mutual sympathies in later years.

McCalman brings the four voyagers' stories together in 1858 when Wallace, who was still in the field at Ternate in the Moluccan Islands (now part of Indonesia), sent Darwin a manuscript articulating a concept very similar to Darwin's as yet unpublished theory of evolution by natural selection. He dwells on Darwin's distressed reaction to this letter, and the quick decision by Hooker and the geologist Charles Lyell to have it read in Wallace's name at the Linnean Society alongside unpublished extracts that proved Darwin's priority. McCalman concludes that this move, though it was "dodgy", helped to bring attention to Wallace's work and provided him with credibility as a theorist and not a mere professional collector. This view of Wallace's rise in prestige feeds into the argument of the book's closing chapters.

To this point, McCalman has relied on the ready metaphor of the individual voyage to describe the personal development of each protagonist. Now he uses the image of a naval fleet to portray how Darwin's three junior colleagues coordinated efforts to support "their admiral's" reputation against enemies inside and outside the scientific community. McCalman wants us to see that their work on behalf of evolutionary theory was part of a larger campaign — by Huxley, Hooker and others such as the physicist John Tyndall — to wrest control of science from the traditional elite. They wanted to put scientific institutions into the hands of middle-class, secular-leaning professionals like themselves and in turn to increase the power of these institutions in Victorian society.

Getting a clear picture of how this post-1859 "battle for the theory of evolution" was fought

and won would require attention to many factors. But McCalman is less concerned with how the battle was waged than with why the participants on Darwin's side felt so strongly about their cause, and this is where he draws connections to the voyages to which he has devoted the bulk of the book.

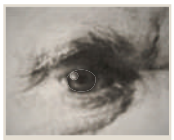
On the intellectual level, each naturalist had his faith in the doctrine of special creation unsettled during his travels by the experience of seeing organisms in unexpected geographical distributions. Then there was the psychological link among those who had earned their knowledge through the alternating peril and drudgery of a long sea voyage. "We have had a masonic bond," Huxley wrote to Hooker in 1888, "in both being well salted in early life." Finally, McCalman's book seems to suggest, we should view the decision of a young Hooker, Huxley or Wallace to undertake such a difficult voyage as the most telling symptom, not the cause, of his enduring devotion to science. ■

Alistair Sponsel is a senior research assistant at the Darwin Correspondence Project, University of Cambridge, UK, and a postdoc in the Department of the History of Science at Harvard University, Cambridge, Massachusetts, USA.

See www.nature.com/darwin/index.html for more on Darwin.

Log of life beneath the waves

World Ocean Census: A Global Survey of Marine Life
by Darlene Trew Crist, Gail Scowcroft and James M. Harding Jr
Firefly Books: 2009. 256 pp. $40, £30



Begun in 2000, the first global marine census is due to be completed next year. In anticipation of the official release of these results, a beautifully illustrated book highlights the findings to date of this massive project. *World Ocean Census* also hints at the studies that might stem from the millions of samples collected.

The Census of Marine Life aims to catalogue the oceans' inhabitants now, place them in a historic context and project what might be found in the oceans of the future. As of 2008, it involved some 2,000 scientists from 82 nations, and had US$500 million of funding. A bold undertaking, it marks a "Herculean decade of exploration", explains oceanographer Sylvia Earle in her foreword to the book.

Science writer Darlene Trew Crist and educators James Harding and Gail Scowcroft emphasize that marine science is in an age of discovery. When the census started, only 250,000 species were known out of the millions estimated to live in the ocean. Researchers expect to find many thousands more, but will undoubtedly fall short of a complete accounting, given the size of the task. Just three recent expeditions to the Southern Ocean yielded some 700 likely new species, and one litre of seawater alone can host 20,000 different microbes.

Like the census, the book is organized by oceans past, present and future. Information dating back 500 years or more is gleaned from old whaling logs, scientific expedition records and even old restaurant menus that provide snapshots of species exploitation. Although not always rigorously quantitative, such records can offer a baseline for conservation targets. Past levels of some exploited fish, such as cod, were surprisingly high. "It is virtually impossible to imagine how much the oceans of the past teemed with life," census researchers have remarked.

Life nevertheless flourishes in today's oceans. The book is at its best when it offers glimpses of the astonishing array of sea creatures revealed by the survey, such as the deepest comb jellyfish ever recorded — found at 7,000 metres — which uses long filaments to anchor itself to the seafloor like a kite. Special sections tell of the widespread loss of bluefin-tuna stocks, the surprisingly long distances travelled by great white sharks and efforts to protect coral reefs.

The book is full of high-quality photographs. These reveal strange, recently discovered species such as the yeti crab, so named for its hairy legs and claws. The text, however, gets bogged down in lengthy and repetitive descriptions of the technologies used to conduct the studies, including tagging techniques and underwater exploration vehicles. Readers may be tempted to skip ahead, wanting to learn what has been found before reading how it was discovered.

Predicting what will live in future oceans is a challenge. Census studies suggest that the loss of exploited fish is coupled to wider shifts in the ecosystem balance, causing problems that will be exacerbated if current fisheries trends persist. Depleted shark populations in the northwest Atlantic, for instance, have led to an overabundance of the cownose ray, a key shark prey that, in turn, wiped out a scallop fishery. On a more positive note, a study based on a review of practices in Norway suggests that Maine lobster fishermen could reduce both the number of traps they use and the length of their fishing season without reducing catch size. This would substantially decrease the number of critically endangered right whales in the North Atlantic killed by entanglement in lobster-fishing gear.

Fittingly, *World Ocean Census* begins and ends with spectacular photos of jellyfish, which are supremely suited to exploiting the niches created by overfishing. How humans respond to the trends revealed by the census will largely determine whether this 'jellification' of the ocean will continue, or if crippled marine populations may have a chance to recover. ■

Mark Schrope is a writer based in Florida, USA.
e-mail: mark@markschrope.com

See www.nature.com/darwin/index.html for the whole biodiversity special.


A. FIFIS/IFREMER

Recent discovery: *Kiwa hirsuta*, named after the goddess of shellfish, is better known as the yeti crab.

C. ARNOLD

Q&A: Bird behaviour, Darwin and dance

Nicky Clayton, a biologist and psychologist who studies the behaviour of birds, and who is also a salsa and tango dancer, collaborated with Rambert Dance Company to create a work commemorating Charles Darwin. As *The Comedy of Change* tours the United Kingdom, she explains how communicating via motion is common to both dance and the natural world.

Why celebrate evolution through dance?
I am fascinated by the showy displays of clever birds, and their extravagant and elegant dances. As Rambert Dance Company's scientific adviser, my challenge was to distil Darwinian ideas to inspire movement, expressive energy and musicality. I did this by combining my knowledge of evolution with my research on the cognitive capacities of corvids — the crows and jays — and my passion for dance.

What does the performance convey?
This beautiful contemporary dance that Mark Baldwin choreographed is foremost a piece of creative art. It doesn't aim to explain Darwin's theories but it may inspire audiences to think about the biology of change, the beauty of the natural world and our place in it.

We focused on three apparently paradoxical features of the biology of change: how similarities within species become differences through variation; how the future becomes the past; and how the natural world can conceal and yet reveal through camouflage and display. These principles allow females to be particular in their choice of males — and also explain why change can be so comical. The inspiration came from watching birds dance.

How did the Rambert dancers respond to these evolutionary ideas?
Artistic director and choreographer Mark Baldwin and I face the same challenge — how to communicate in the absence of language. To do so, we rely on movements, expressive energy and the dynamics of behavioural displays. The main difference is that my subjects have feathers and no hands. I gave a lecture about Darwinian ideas, and Mark and I did some tango moves afterwards in the studio with the dancers. We conveyed examples of birds dancing, from the synchronized movements of the rooks to the mating dances of male blue manakins and birds of paradise. I also showed them the camouflage displays of octopus, and how one side of the octopus can signal 'keep out' while the other flirts with a female. I drew a number of analogies with dance. For example, as a follower, I could be dancing tango with one man while another creeps up behind to steal me away — an example of male–male competition in sexual selection if ever I saw one.

H. GLENDINNING/RAMBERT DANCE COMPANY

Which birds inspired you most?
One of the most impressive is the male blue manakin, which can be found in the forests of Argentina. It spends 90% of its waking day, for most of the year, in a dancing duel. Its dancing skills matter: females mate only with the best and most expert male dancers. I call it an 'avian tango'; in Argentina, tango was originally danced by pairs of men, and similarly, in blue manakins, an expert leader and an apprentice follower perform a male double act.

Practice is critical: the young birds start when they are two years old, but they do not become an apprentice until they are about eight. Only when the senior male dies may they become the leader.

Despite the male bird's colourful plumage, he is easily concealed in the forest undergrowth. It is only when he dances that his true colours are revealed.

Even more impressive are the western parotia males, which have an 'inflatable tutu' that they use when wooing the ladies, along with the most amazing head and neck isolation movements. In this case of 'now you see it, now you don't', the idea of conceal and reveal comes to the fore. The birds also need years of experience to become the principal male.

How important is coordinated movement in the natural world?
It is very important. You see synchronicity, connection and coordination in the movements of a pair of corvids. These birds tend to pair for life, and maintaining relationships cannot be easy.

I wonder whether these synchronous movements help to cement the pair bond between the two birds, signalling commitment, connection and coordination. This might be especially important in long-lived, large-brained animals that practise long-term serial monogamy — which is true of both the corvids and many human societies.

Do birds respond to music?
Yes, sound and vision are crucial in the bird world, from the showy dancing displays of the birds of paradise to the melodious songs and rich composition of the lyrebird. So it comes as no surprise that vocal-learning birds, which have their sensory apparatus attuned, have provided examples of song learning and of dancing in time to human music.

Three famous case studies of dancing to the beat are Alex the African grey parrot and cockatoos called Snowball and Frostie. And next time you hear the third movement of Mozart's Piano Concerto No. 17 in G, K. 453, think of the composer and his pet starling: Mozart taught the bird to sing the opening theme. ■

Interview by **Patrick Goymer**, associate editor at *Nature*.

See www.nature.com/darwin/index.html for more on Darwin.

In Retrospect: The earliest picture of evolution?

Ideas about the mutability of species may have been part of Enlightenment imagery before Lamarck.

De anima brutorum commentaria (Commentary on the Soul of Animals)
by Francesco Maria Soldini
Gaetano Cambiagi: 1776

Illustrations can convey scientific ideas as effectively as the written word. It is widely accepted that there was no graphical representation of the evolution of species before 1800, when the naturalist Jean-Baptiste Lamarck added the axis of time to his classification tree diagrams. However, I have found illustrations in a little-known book written by the Carmelite monk Francesco Maria Soldini that predate Lamarck's imagery. Plates in the book, published in Florence in 1776, clearly depict life emerging from the sea onto land.

Written in Latin, *De anima brutorum commentaria* is one of many books printed in the sixteenth and eighteenth centuries on the concept that animals have a soul. Soldini anchors his arguments to the writings of great philosophers from ancient times, such as Aristotle, and to scripture, especially the Book of Genesis. He was also influenced by contemporaries, notably Immanuel Kant, Gottfried Leibniz, Étienne Bonnot de Condillac and Pierre-Louis Moreau de Maupertuis. The book is a very rare volume and its scientific significance has escaped notice until now.

Soldini's book is embellished with eight stunning engravings by an unknown artist that depict natural scenes and animals, bound at the beginning of each chapter. The plates are printed in blue or red to contrast with the elegantly decorated initial letter of the text. But the iconographical content of the images is independent from the writing: many of the animal pictures are taken from the wood carvings of other treatises on zoology, such as those by the sixteenth-century naturalists Ulisse Aldrovandi and Conrad Gesner, which are not cited by Soldini. It is therefore likely that the plates were added separately to decorate the book, which was produced by Gaetano Cambiagi, typographer to the Grand Duke of Tuscany.

In two of the eight plates, the engraver portrays marine animals, mainly crustaceans, leaving the water and colonizing land. These images are reminiscent of the 'Neptunian' theory of Earth described by the French literary scholar and diplomat Benoît de Maillet (1656–1738). In his book *Telliamed*, which circulated for decades before being posthumously published in 1748, de Maillet explains how Earth was once entirely covered by water. He proposes that life began in the water in the form of minuscule seeds that joined together to create all aquatic forms, from which all terrestrial and winged creatures were then derived. In his opinion, all plants and animals would have analogous examples among the aquatic specimens. The plates do indeed display marine animals that have parallels with species living on land, even in their names, such as the mantis shrimp and the fantastical marine rhinoceros. Yet Soldini makes no mention of either *Telliamed* or de Maillet in his otherwise highly referenced book.

A third plate represents images of fish taken from the iconography of the sixteenth century. They swim holding their heads above water with birds flying above. Such representations follow another de Maillet idea, that animals are derived from two basic types: the flying ones that live between the sea floor and the surface, which today is known as the pelagic zone, and the creatures that crawl on the sea floor, or benthic zone. Birds would have stemmed from the flying type, terrestrial animals from crawling forms.

The anonymous plates in *De anima brutorum commentaria* demonstrate the extent to which evolutionary ideas circulated during the Enlightenment, when drawing and carving were valuable means of transmitting progressive ideas to readers with minds open to novel concepts. ■

An engraving published in 1776, 83 years before *On the Origin of Species*.

Fausto Barbagli is curator at the Natural History Museum of Florence University, Via Romana 17, I-50125 Florence, Italy.
e-mail: fausto.barbagli@unifi.it

See www.nature.com/darwin/index.html for more on Darwin.

Corrections
In the Book Review 'Amphibian mystery misread' by Alan Pounds and Karen Masters (*Nature* **462,** 38–39; 2009) the sentence "Collins and Crump's selection of published work and quoted opinions downplays such links" should have read "Collins and Crump's assessment of published work and the opinions that they quote downplay such links."

In Alison Abbott's Arts Review 'Florence's observatory restored' (*Nature* **462,** 40; 2009), "Pietro Leopardo" should have read "Pietro Leopoldo".

NEWS & VIEWS

ASTROPHYSICS

Burst of support for relativity

Giovanni Amelino-Camelia

Light from a distant γ-ray burst backs up a key prediction of Albert Einstein's theory of relativity — that photon speed is the same regardless of energy. But it might set the stage for evolution of the theory.

The concept of relativistic theory — a formulation of the laws of physics that is independent of the state of motion of observers — has enjoyed remarkable longevity. But it has not been a smooth ride. Nearly all major triumphs of physics challenged the concept and ultimately required its revision. The first relativistic theory, Galilean relativity, was a key element of Newtonian mechanics, but became evidently inadequate when a satisfactory understanding of electromagnetism matured. That took relativity out of fashion for a while.

It was Albert Einstein's theory of special relativity that later led to its re-emergence in a form[1] encoding the novel principle that the speed of light is the same for all observers in uniform relative motion. And when Isaac Newton's law of gravitation was replaced by Einstein's more accurate description of gravitational phenomena, on the basis of the understanding that space bends in response to the energy carried by particles, the process ended up requiring a significant generalization of special relativity. This led to Einstein's theory of general relativity, still in use today, which established that the laws of physics are independent of the state of motion of observers, even when these motions are non-uniform. On page 331 of this issue, Abdo *et al.*[2] put general relativity under scrutiny by analysing observations of a distant and short-lived γ-ray burst, dubbed GRB 090510, undertaken with the Fermi Gamma-ray Space Telescope.

The motivation for testing general relativity comes indirectly from quantum mechanics, perhaps the greatest achievement of physics. It has taught us how most of the physical properties formulated by Newton in terms of continuous variables are more accurately described in terms of finite small bits called 'quanta' (Fig. 1). In the ongoing process of learning how to apply this 'quantum paradigm' to the description of space-time[3] — a central aspect of the quantum gravity problem[4] — we might uncover profound implications for general relativity. In particular, several theories predict that tiny space-time quanta[5] can affect the speed of photons (the particles that make up light) by inducing a correspondingly small dependence on photon energy — a feature that is not allowed in general relativity.

AKG-IMAGES

Figure 1 | Quanta and pointillism. The mechanism behind the workings of quantum mechanics is not too different from the one used in the style of painting called pointillism — here depicted in Georges Seurat's painting *La Seine à la Grande Jatte*. In this form of painting, the low resolution of our eyes, particularly when looking across a relatively large distance, is exploited to generate the impression of a wide selection of colours and blending when, in fact, the painter uses many small distinct dots of only a few colours. Similarly, quantum mechanics describes physical properties in terms of tiny bits called quanta. Whether such a description is valid for space-time is an open question. Abdo and colleagues[2] analyse observations of a distant γ-ray burst to look for indirect evidence of quanta of space-time.

For the law of quantum space-time that fixes the dependence of the speed (v) of photons on photon energy (E), several arguments favour a formulation of the type $v = c\,(1 - \eta E^n/E_P{}^n)$, where c is Einstein's maximum-velocity scale, η is a parameter to be determined experimentally, n is a model-dependent integer and E_P is the energy scale known as the Planck scale. The large value of the Planck scale, about 10^{28} electronvolts, reflects our best current estimates of the ultra-small length scale that characterizes space-time quanta. For the integer n, different models are found to give either $n = 1$ or $n = 2$, and this is not a minor difference: because of the large value of the Planck scale, the energy dependence of photon speed is already extremely weak in the case $n = 1$, and for $n = 2$ the effect is even smaller.

The Fermi Gamma-ray Space Telescope is ideally suited to look for traces of the energy-dependence of photon speed in γ-ray bursts — intense bursts of γ-ray photons emitted from sources at large cosmological distances from Earth. But before the telescope's launch, it was difficult to estimate how accurately it would allow such an investigation to be made. Like all observatories (in contrast to controlled laboratory experiments), Fermi's effectiveness

depends strongly on what it happens to catch from the sky. The γ-ray burst GRB 090510 was a great catch on various grounds[2], including its short duration and the presence of several high-energy photons.

In their study, Abdo and colleagues[2] find no evidence for a dependence of photon speed on energy. But this is one of those rare instances in which a negative result still generates much scientific excitement. In fact, using my previous notation, the limit set by the authors' analysis amounts to η of less than about 0.8 for the case with $n = 1$. Before the advent of the Fermi telescope, we could only probe values of η that were significantly bigger than 1, yet the models of quantum space-time suggest that η should be roughly of the order of 1.

For relativity traditionalists, the authors' result is big news, especially in light of a preliminary analysis of γ-ray-emission data from an active galactic nucleus reported in the past year[6]. These data, obtained with the ground-based MAGIC telescope at La Palma in the Canary Islands, were tentatively interpreted in support of violations of general relativity (values of η between 6 and 17). Fermi's observation of GRB 090510 rules out such an interpretation, and seems to set the stage for even more detailed confirmations of the current formulation of relativity. But there is also something for those who would rather witness another reformulation of relativity: now that Fermi has demonstrated its ability to probe how photon speed depends on energy, it is conceivable that one of its next γ-ray-burst observations will actually provide evidence for the much-sought-after energy dependence.

As usual in science, the smart money is on the most conservative expectations. Nature, with its unique clever ways, might have figured out how to quantize space-time without affecting relativity. Or perhaps the effect is significantly weaker, only appearing at the level $n = 2$, so that it will remain hidden from us until we have telescopes that are suitable for studying the analogous effects of space-time quantization for the ultra-high-energy neutrinos that accompany γ-ray bursts[7]. For the field of relativity and space-time physics, which stood still for nearly a century while looking in on the tremendous discoveries produced by the quantum paradigm in other areas of physics, even a slim chance of being on the verge of a new revolution is truly exciting. ■

Giovanni Amelino-Camelia is in the Department of Physics, Università di Roma La Sapienza, and the Istituto Nazionale di Fisica Nucleare (INFN), Sezione Roma 1, Piazzale Moro 2, Rome 00185, Italy.
e-mail: giovanni.amelino-camelia@roma1.infn.it

1. Stachel, J. *Nature* **433,** 215–217 (2005).
2. Abdo, A. A. *et al. Nature* **462,** 331–334 (2009).
3. Amelino-Camelia, G. *Nature* **398,** 216–218 (1999).
4. Carlip, S. *Rep. Prog. Phys.* **64,** 885–942 (2001).
5. Amelino-Camelia, G. *Nature* **418,** 34–35 (2002).
6. Albert, J. *et al. Phys. Lett. B* **668,** 253–257 (2008).
7. Jacob, U. & Piran, T. *Nature Phys.* **3,** 87–90 (2007).

STRUCTURAL BIOLOGY

New beginnings for transcription

Steven Hahn

A structure for the enzyme RNA polymerase II in combination with the transcription factor TFIIB changes our view of how the polymerase and its helper proteins initiate transcription.

In all organisms, the initiation of transcription — RNA synthesis from a DNA template — is a collaborative process involving cooperation between RNA polymerase (Pol), the enzyme that transcribes RNA from DNA, and general transcription factors that assist initiation in various ways. Tremendous progress has been made over the past 20 years in understanding how transcription in eukaryotes (plants and animals) begins, including the elucidation of structures of several transcription initiation factors and of Pol II — one of the eukaryotic polymerases[1–3]. What's been missing, however, is the structure of Pol II combined with general factors and its DNA substrate. This is partly because this initiation complex is held together by many weak protein–protein and protein–DNA interactions, and it is difficult to assemble it in large enough quantities for structural analysis.

On page 323 of this issue, a breakthrough paper by Kostrewa *et al.*[4] reports the structure of the yeast general transcription factor TFIIB in complex with Pol II, providing unprecedented insight into the configuration of the transcription-initiation machinery. This work overturns several key elements of a previous Pol II–TFIIB structure[5], altering our view of the mechanism by which TFIIB functions in initiation.

All eukaryotes have at least three nuclear RNA polymerase enzymes (Pol I–III), which synthesize different classes of RNA. Prokaryotes (bacteria and archaea) have only one polymerase. But despite these differences, all of these multi-subunit polymerases use the same general mechanism to initiate transcription. First, they take up position at specific DNA sequences termed promoters, which determine the start site of transcription, in a protein–DNA complex made up of the polymerase itself and one or more general transcription factors. This forms the preinitiation complex (PIC), or closed complex. Next, about 10 bases of the DNA double helix separate, and the single-stranded DNA template strand slips into the active site, which lies in a deep cleft in the polymerase enzyme, to form the open complex state.

Initial synthesis of RNA by the polymerase is problematic, generating many short abortive RNA products. This is partly due to the low stability of short DNA:RNA hybrids in the enzyme active site. Once an RNA 7–10 bases long is synthesized, the polymerase releases its contacts with the promoter and initiation factors to enter a processive elongation form termed the elongation complex[3], which synthesizes full-length RNA transcripts. The polymerase performs these actions with the assistance of essential transcription factors: σ factor for bacterial Pol and general transcription factors for archaeal and eukaryotic Pol enzymes. These factors perform various tasks, aiding recruitment of the polymerase by activators, recognition of specific DNA sequences near the transcription start site, DNA unwinding, and stabilization of the DNA:RNA hybrid.

TFIIB and TATA binding protein (TBP) are the two general transcription factors that hold the key to determining the architecture of the Pol II PIC. The two TFIIB structural domains (called the zinc-ribbon and the core domain) bind to Pol II at separate locations[5,6], with the flexible polypeptide that joins these two domains positioned within the Pol II active-site cleft, where its functions are to help form the open complex and recognize the transcription start site. The TFIIB core domain also binds to TBP; TBP recognizes the TATA sequence found in some promoters and bends promoter DNA around the Pol II surface and over the cleft.

A previous low-resolution Pol II–TFIIB structure[5] had suggested that the TFIIB core domain binds Pol II far from the cleft and that the flexible region joining the TFIIB zinc-ribbon and core domains forms a hairpin structure termed the B-finger, which was proposed to recognize DNA sequences near the transcription start site. Kostrewa and colleagues' new structure[4] contradicts both of these proposals. The authors find that the TFIIB core domain lies close to the Pol II cleft, in agreement with previous biochemical studies[6,7]. They also find that the TFIIB flexible region, formerly called the B-finger, forms a helix and a flexible loop that they call the reader region (Fig. 1a, blue) because it is predicted to be in a position that can 'read' the DNA sequence eight bases upstream of the transcription start site — a position evolutionarily conserved in promoters of the yeast *Saccharomyces cerevisiae*. Consistent with this hypothesis, mutations in the yeast TFIIB reader region are known to cause transcription initiation downstream of the normal start site, as if specificity has been lost or altered in the mutants[1].

Kostrewa and colleagues then merged their PIC model with the known structure of the transcription elongation complex[3] to reveal the first detailed model for the structure of the open complex (Fig. 1b). The authors[4] compared the predicted path taken by DNA in the PIC and open complex, showing that

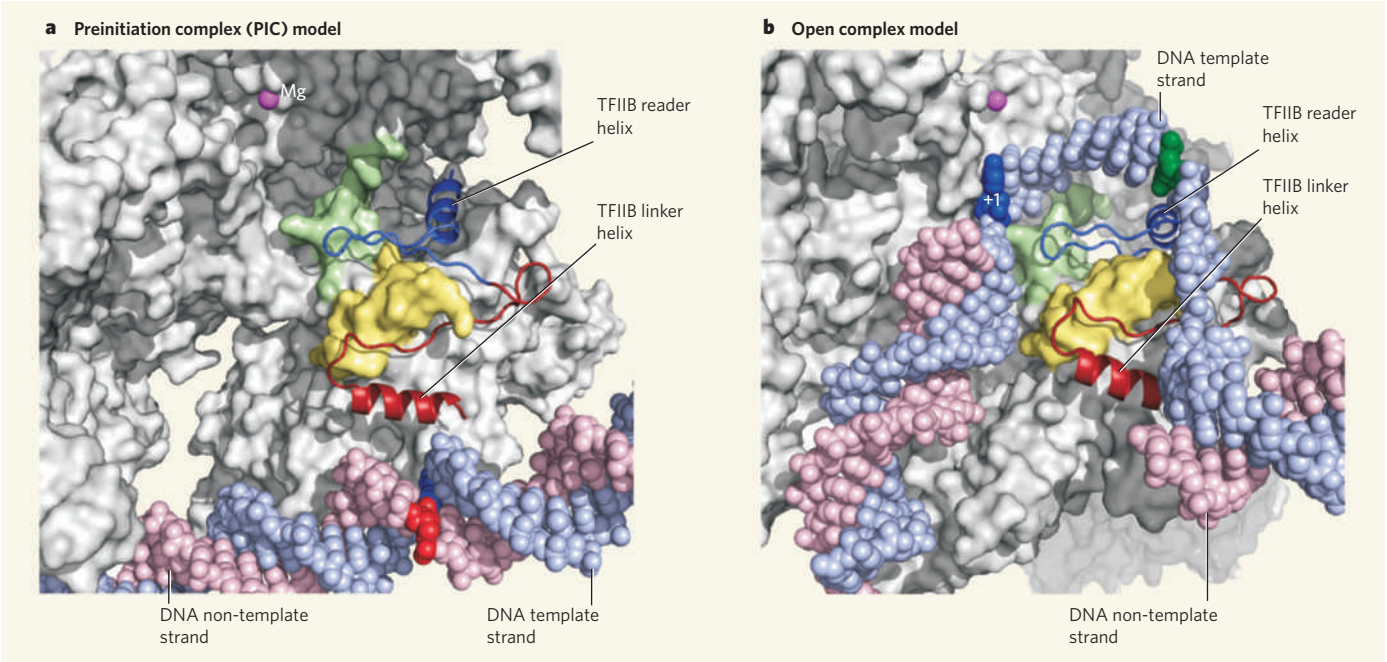


Figure 1 | Structures of RNA polymerase II in combination with the general transcription factor TFIIB. **a**, Cutaway view of the active-site cleft of polymerase II (Pol II) in Kostrewa and colleagues' model[4] of the preinitiation complex (PIC), showing the position of the active-site magnesium atom (magenta), the TFIIB flexible region (TFIIB reader helix, blue; TFIIB linker helix, red) and promoter DNA (template strand, light blue; non-template strand, light pink), with the farthest upstream DNA base pair to be 'melted' coloured dark blue and red. The surface of the Pol II large subunit is shown, with the flexible region termed switch 2, which is involved in DNA opening, coloured green, and the rudder domain, involved in initiation and elongation, coloured yellow[2,3]. **b**, Cutaway view of the model of the Pol II open complex with the same colour scheme as above, except that the template base at the transcription start site (+1) is dark blue. Note that the reader helix is adjacent to position –8 (green) on the template strand and the linker helix is positioned near the site of DNA-strand separation.

the second flexible TFIIB region, termed the linker, forms a strand, loop and helix located between the double-stranded DNA in the PIC (Fig. 1a) and the single-stranded DNA in the open complex (Fig. 1b). This suggests that the TFIIB linker is involved in opening and/or stabilization of the DNA 'bubble'.

To confirm this, the authors examined the archaeal polymerase enzyme. Archaea have an elegantly simple transcription system that uses only three general initiation factors, corresponding to TBP, TFIIB (TFB) and TFIIE (TFE). The B-reader and linker are evolutionarily conserved in archaeal TFB, and TFB proteins that have mutations in the linker region are defective in DNA-strand separation, as are archaeal polymerase enzymes that have mutations in the clamp domain, which is predicted to bind and position the TFB linker helix. The effects of these TFB and Pol mutations can all be suppressed by pre-opening the DNA double helix[4], showing that opening of the helix is the principal function of the TFB linker region.

Although the sequences of bacterial σ factor and TFIIB are very different, both factors position a flexible polypeptide region within the polymerase active site[8,9], albeit with different amino- and carboxy-terminus polarity. Thus, multi-subunit polymerases share a common requirement for an accessory factor within their active site to promote initiation. As these factors block the RNA exit channel (from which the newly transcribed RNA emerges), they contribute directly to abortive RNA synthesis, and must be ejected from the active site for transcription to begin[8]. It is thought that σ initiates DNA-strand separation by competing with base pairing at the site where the DNA strands separate. Although the eukaryotic Pol II system needs additional factors, including a DNA helicase, for DNA opening, TFB seems to be sufficient to open DNA in the archaeal system, suggesting that the archaeal TFB and the eukaryotic TFIIB linker regions are directly involved in DNA-strand separation[4]. In Kostrewa and colleagues' new structure[4], the TFIIB linker lies adjacent to the junction of single- and double-stranded DNA (Fig. 1b).

With this new information, many questions arise: how strongly conserved is the function of the TFIIB reader and linker? Pol I is the only eukaryotic Pol that does not use a TFIIB-like factor, but presumably the function of the TFIIB flexible region is fulfilled by another Pol I general factor with a sequence that is too highly divergent to be identified by sequence comparison. The Pol III factor Brf (TFIIB-related factor) contains conserved zinc-ribbon and core domains, but the sequence in the flexible region joining these domains is highly divergent. However, Brfs have conserved blocks of sequence in nearly the same positions as the B-reader helix, strand and the linker helix, and they probably fulfil the same functions in Pol III initiation. Although the flexible region is found in most archaea, there are some interesting exceptions[10]. Some archaea encode multiple TFB factors that are differentially regulated or that associate with the polymerase at different genes. For example, *Pyrococcus furiosus* encodes Tfb1 and Tfb2, the latter having a completely different B-reader helix and also lacking the B-reader linker strand and loop. One explanation for this difference is that the distinct sequence of the reader helix allows recognition of alternative promoter sequences, contributing to gene-specific regulation in archaea.

The landmark results from Kostrewa and colleagues' TFIIB–Pol II structure and functional analysis[4] set the stage not only for a deeper understanding of transcription initiation by all multi-subunit Pol enzymes, but also for understanding how the fundamental mechanism of transcription has been evolutionarily conserved among all organisms. ■

Steven Hahn is at the Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA.
e-mail: shahn@fhcrc.org

1. Hahn, S. *Nature Struct. Mol. Biol.* **11,** 394–403 (2004).
2. Cramer, P., Bushnell, D. A. & Kornberg, R. D. *Science* **292,** 1863–1876 (2001).
3. Westover, K. D., Bushnell, D. A. & Kornberg, R. D. *Science* **303,** 1014–1016 (2004).
4. Kostrewa, D. *et al. Nature* **462,** 323–330 (2009).
5. Bushnell, D. A., Westover, K. D., Davis, R. E. & Kornberg, R. D. *Science* **303,** 983–988 (2004).
6. Chen, H.-T. & Hahn, S. *Cell* **119,** 169–180 (2004).
7. Chen, H. T., Warfield, L. & Hahn, S. *Nature Struct. Mol. Biol.* **14,** 696–703 (2007).
8. Murakami, K. S., Masuda, S. & Darst, S. A. *Science* **296,** 1280–1284 (2002).
9. Vassylyev, D. G. *et al. Nature* **417,** 712–719 (2002).
10. Micorescu, M. *et al. J. Bacteriol.* **190,** 157–167 (2008).

EVOLUTIONARY BIOLOGY

Why reproduction often takes two

Aneil F. Agrawal

On the face of it, self-fertilization is the efficient way to breed: compared with outcrossing, there's usually much less fuss, for a start. So why isn't reproduction by selfing far more prevalent than it is?

Finding a mate can be difficult or even dangerous. Moreover, in many organisms one parent — typically the female — makes a much larger energetic contribution to offspring than the other parent, even though offspring receive an equal number of genes from both parents. Because of these costs, uni-parental forms of reproduction, such as self-fertilization and asexual reproduction, would seem a more evolutionarily sensible alternative[1,2]. Although many plants and animals are capable of self-fertilization, most species reproduce completely or partially through bi-parental reproduction or outcrossing[3,4]. The prevalence of outcrossing implies that there are advantages to this mode of reproduction that outweigh its substantial costs.

Among the hypothesized benefits of outcrossing are that it reduces the effects of deleterious mutations, or that it improves the ability to adapt to changing environmental circumstances. Relevant data that bear on these possibilities have been limited in scope. In this issue (page 350), however, Morran *et al.*[5] report an experimental study that demonstrates the existence of both advantages of outcrossing.

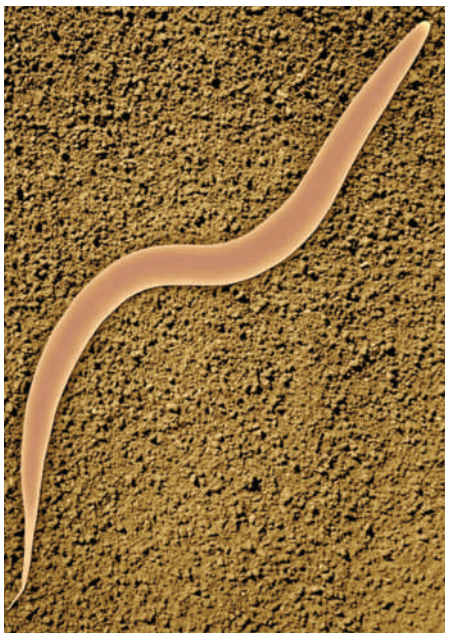
All populations harbour variants (alleles) of genes that are deleterious because mutation is constantly converting good alleles into bad ones ($A \rightarrow a$). Such mutations are often recessive, meaning that most of the negative effects are masked in heterozygotes (*Aa*). Self-fertilization puts these alleles into the homozygous (*aa*) state, thereby exposing them to selection and causing a reduction in fitness known as inbreeding depression. A classic explanation for the maintenance of outcrossing is the avoidance of inbreeding depression. However, when populations reproduce by selfing, recessive deleterious mutations are expressed and eliminated by selection, reducing the magnitude of future inbreeding depression. Consequently, the act of self-fertilization can make it easier for selfing to evolve in subsequent generations[6,7].

But inbreeding depression can never be completely purged, for two reasons. First, selection is ineffective at eradicating mutations of small effect, especially in smaller populations. Second, new mutations are constantly occurring. If deleterious alleles of small effect are introduced at a sufficiently high rate, then inbreeding depression will be maintained at a sufficiently high level to provide an ongoing advantage to outcrossing.

Outcrossing is also hypothesized to be advantageous by allowing for faster adaptation[8,9]. In highly selfing populations, each beneficial mutation is more or less trapped on the genetic background on which it arose. This reduces the rate of adaptation for two reasons. First, the genetic background may contain deleterious alleles at other loci, thus hampering the spread of the beneficial mutation. Second, by being trapped in its original genotype, the beneficial mutation cannot combine with other new beneficial mutations that may have occurred on other genetic backgrounds.

By contrast, when populations are outcrossing, a process of genetic recombination between different genomes occurs, which allows a new beneficial mutation to escape deleterious alleles on its original background and to combine with other beneficial alleles that arise elsewhere in the population. In selfing populations, individuals are largely homozygous, and recombination has no effect on the distribution of alleles, even though genetic crossing-over occurs. Outcrossing is advantageous because it puts different sets of chromosomes together, allowing crossing-over to result in meaningful genetic exchange.

These, then, are two potential advantages of outcrossing. Morran *et al.*[5] tested them with experiments involving populations of the nematode *Caenorhabditis elegans* (Fig. 1). This worm normally exhibits a low to moderate level of outcrossing. But by making use of some well-studied mutants, Morran *et al.* were able to create populations that were either obligately outcrossing or obligately selfing. They then subjected these populations to various evolutionary conditions to investigate how the degree of selfing affects evolutionary outcomes.



DENNIS KUNKEL MICROSCOPY, INC./VISUALS UNLIMITED/CORBIS

Figure 1 | Fit for evolutionary studies — the nematode *Caenorhabditis elegans*. When it comes to reproduction, one will do, but Morran *et al.*[5] find that outcrossing has evident advantages over selfing (×150).

In the first experiment, populations were repeatedly exposed to a mutagen, artificially inflating the mutation rate to about twice the normal rate. So that more mutations would affect fitness, these populations were kept in a challenging environment that required the worms to traverse a rugged terrain, as opposed to the benign conditions typically used for rearing *C. elegans* in the lab. After 50 generations, the highly selfing populations showed substantial declines in fitness, as predicted if deleterious alleles cannot be efficiently eliminated. By contrast, obligately outcrossing populations showed no decline.

In a parallel treatment, another set of populations was allowed to evolve in the same conditions but at the normal mutation rate. At this lower mutation rate, the outcrossing populations improved in fitness over time, adapting to the rugged terrain. But the highly selfing populations showed no evidence of adaptation. To further test whether outcrossing facilitates adaptation, Morran *et al.* performed another experiment in which a set of populations evolved in environments containing a virulent bacterial pathogen. After 40 generations, the obligately outcrossing populations showed a staggering 150% increase in fitness, whereas the obligately selfing populations did not adapt at all.

These experimental studies[5] demonstrate a clear advantage of outcrossing populations over selfing populations, and they parallel previous work showing that sexual populations have advantages over asexual populations with respect to the rate of adaptation[10,11]. Although bi-parentally reproducing populations may be more fit, this does not guarantee that alleles causing uni-parental reproduction will not easily invade a population — the individual-level benefits of selfing may be much more important than the group-level costs. The work of Morran *et al.*[5] is particularly intriguing because the authors found evidence that outcrossing evolved within those of their wild-type populations that were not genetically constrained to be obligately selfing or obligately outcrossing. Outcrossing rate in wild-type populations increased in populations subjected to elevated mutation rates, and possibly also in populations adapting to the pathogen.

Experimental evolution studies such as these provide a fundamental first test of theory. However, more detailed experiments are needed to provide a clearer understanding of why reproductive strategies evolve within such populations; are short-term or long-term effects responsible for this type of evolution[12]? By being amenable to a variety of coarse- and

fine-scale approaches, such experimental systems hold great promise for helping us to achieve a better understanding of the patterns we see in nature. ■

Aneil F. Agrawal is in the Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario M5S 3B2, Canada.
e-mail: a.agrawal@utoronto.ca

1. Fisher, R. A. *Ann. Eugen.* **11,** 53–63 (1941).
2. Lively, C. M. & Lloyd, D. G. *Am. Nat.* **135,** 489–500 (1990).
3. Goodwillie, C., Kalisz, S. & Eckert, C. G. *Annu. Rev. Ecol. Syst.* **36,** 47–79 (2005).
4. Jarne, P. & Auld, J. R. *Evolution* **60,** 1816–1824 (2006).
5. Morran, L. T., Parmenter, M. D. & Phillips, P. C. *Nature* **462,** 350–352 (2009).
6. Lande, R. & Schemske, D. W. *Evolution* **39,** 24–40 (1985).
7. Charlesworth, D., Morgan, M. T. & Charlesworth, B. *Evolution* **44,** 1469–1489 (1990).
8. Haldane, J. B. S. *The Causes of Evolution* (Harper, 1932; reprinted with introduction and afterword by E. G. Leigh, Princeton Univ. Press, 1990).
9. Peck, J. *Genetics* **137,** 597–606 (1994).
10. Colegrave, N. *Nature* **420,** 664–666 (2002).
11. Goddard, M. R., Godfray, H. C. J. & Burt, A. *Nature* **434,** 636–640 (2005).
12. Agrawal, A. F. *Curr. Biol.* **16,** R696–R704 (2006).

PALAEOCLIMATE

Kink in the thermometer

David Noone

Temperature estimates derived from isotopes in polar ice cores reveal much about Earth's past climate. According to the latest analysis, interglacial periods were rather warmer than previously thought.

For the past million years or so, the transition between glaciations and warm interglacials has been dominated by a cycle of about 100,000 years. The last warm period, the Eemian, occurred around 128,000 years ago, and from various proxy measurements it is widely accepted that temperatures then were higher than those during modern pre-industrial times. The Eemian therefore offers a window on the consequences of contemporary global warming. For instance, the warmer climate was associated with significant changes in the volume of the Greenland ice sheet, and as a result sea level could have been about 4–6 metres higher than it is today[1].

In this issue, Sime and collaborators[2] (page 342) describe how they have used climate-model simulations to re-examine the temperature history recorded by the isotopic composition of three long ice cores from Antarctica. The authors conclude that previous estimates of southern-polar warmth during the Eemian and other interglacials have in fact been underestimates. They suggest that temperatures peaked at some 6 kelvin higher than at present, about double the usually accepted figure.

Climate records from polar ice cores provide arguably the best indication of global environmental change on timescales from centuries to many millennia. Like tree rings that mark each year by a season of growing, the continuous accumulation of snow on the ice sheets provides a chronology. The isotopic chemistry of the ice then provides a measure of climate at the time the snow fell. This use of the stable hydrogen- and oxygen-isotope composition of polar snow as an indicator of past temperature comes from pioneering work performed almost half a century ago[3,4]. At sites outside the tropics, a strong relationship exists between annual mean temperature and the annual mean isotopic composition of precipitation[5], which hints at the possibility of using the isotopes from a single site as a palaeothermometer. However, the question remains as to what degree this spatial information can be used to interpret the temporal variability captured in polar ice cores[6,7].

The mechanism leading to the temperature–isotope relationship at high latitudes is the preferential removal of the heavy nuclides oxygen-18 and deuterium (^{18}O and ^{2}H) during precipitation. During transport to the cold polar atmosphere, continual precipitation from an air mass originating at a lower latitude will deplete that air mass in the heavy nuclides. The isotopic composition will depend on the fraction of the original water mass remaining, which depends exponentially on temperature via the Clausius–Clapeyron relation. This gives an approximately linear relationship between the isotopic composition of precipitation and temperature, with a slope of about 8‰ per kelvin for ^{2}H (Fig. 1). Knowing this slope, one can in principle convert measured changes in isotopic composition to a temperature scale.

There are, however, various confounding factors[8]. The isotopic composition of vapour at the source region is unlikely to remain constant; and changes in the distribution of the moisture source will change the isotope–temperature slope, because less-distant sources are less depleted and lead to a reduction of the slope[9]. Further, different cloud microphysical processes dictate a different efficiency in isotopic fractionation. Also, the average isotopic composition of annual layers of snow will be biased towards that of the season in which most of the precipitation falls, so changes in the time of year when snow falls could distort the reconstruction of annual mean temperature. Such factors mean that the spatial relationship

50 YEARS AGO

It has often been suggested that the dark areas of Mars consist partly of vegetation, particularly in view of the seasonal variation of the intensity of the dark regions. Tests for the high near-infra-red reflectivity characteristic of many plants have all given negative results. A few terrestrial plants, such as some lichens, do not show this characteristic, and possibly such plants are present on Mars. W. M. Sinton ... has suggested and twice carried out a new test for the presence of vegetation. All organic molecules possess strong absorption bands at wave-lengths near 3.4μ ... The radiation received from Mars was analysed theoretically into thermal radiation and reflected solar radiation. The latter shows three absorption bands at 3.43μ, 3.56μ and 3.67μ ... Although one cannot be certain that no inorganic molecule can explain these absorption bands, the observed spectrum does fit very closely that of organic compounds and plants ... Sinton's results are the best evidence yet produced for the existence of vegetation on Mars.

From *Nature* 21 November 1959.

100 YEARS AGO

With reference to the recent paper by Dr. Pocklington before the Royal Society, on the functions of the Martian canals ... I should like to suggest that these canals may perhaps be used for power-storage purposes. In Mars, possibly, there are seasons of winds or monsoons during which the upper reaches of the canals would be pumped full by innumerable windmills, and the power thus stored utilised during calm seasons, and transmitted electrically for lighting, heating, and general power purposes. For a population which had exhausted all its mineral fuel, which possessed no extensive ocean, and whose soil and climate were unsuitable for the growth of fuel, this would indeed appear to be the only means of obtaining heat and power.

From *Nature* 18 November 1909.

50 & 100 YEARS AGO

NATURE|Vol 462|19 November 2009

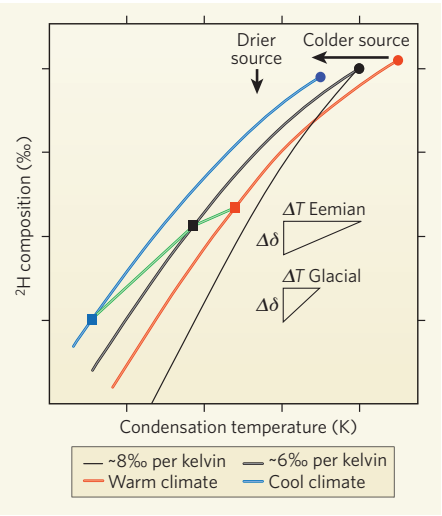


Figure 1 | Temperature–isotope relationships. As a moist air mass moves polewards from a lower-latitude source, and experiences continuous precipitation, the preferential loss of the heavy, ^{2}H, isotope gives an approximately linear dependence of ^{2}H composition on temperature. The result that emerges in spatial data[6] is shown by the thin black line (slope ~8‰ per kelvin). Additional mixing during transport to the pole reduces the slope (thick black line, ~6‰ per kelvin). Change in the mean temperature and humidity of the source region (dots) will change the position of the curve but not the slope (blue and red lines). Change in the isotopic composition for a fixed ice-core site (squares) combines the spatial slope with any changes in the air-mass origin and mixing, and in cloud microphysics. The green lines depict Sime and colleagues' conclusion[2]: a reduced slope for warmer climates such as the Eemian implies a larger temperature change (ΔT) for a given isotopic difference ($\Delta\delta$) compared with the slope for glacial climates.

between the isotopic composition and temperature is more typically about 6‰ per kelvin. For reliable reconstructions, the extent to which these atmospheric processes influence the temporal slope must come into question — in fact, the slope itself can perhaps be seen as a measure of the processes.

In their re-evaluation of the evidence, Sime *et al.*[2] used isotope records from three Antarctic ice cores extending back to 340,000 years ago. These data show that the relevant temporal isotope slope differs from location to location, and that the slope is generally smaller during times when the climate is warmer. In their model simulations, the authors tested whether the isotope changes were attributable to local temperature change or to changes in the isotope–temperature slope. Their results show that the latter dominates. While challenging a tenet of temperature reconstructions from isotopes, their findings imply that the processes that control regional hydrology need not directly control the local annual mean temperature. During cold climates, the relatively small changes in the slope are consistent with temperatures inferred from earlier studies[6,10]. But the lower slope during warmer, interglacial climates leads Sime *et al.*[2] to conclude that temperatures were higher than those derived using the usual slope (Fig. 1).

The idea that interglacial climates were warmer than previously thought raises questions about the strength of climate feedbacks and regional amplification of warming at high latitudes. It is also relevant to the question of the stability of the existing Greenland and West Antarctic ice sheets — at what level of warming will large-scale melting occur?

The principle that the relationship between isotope composition and temperature is non-linear represents a coming-of-age for reconstructions of this kind: it is clear that a simple relationship is not adequate to capture the influence of atmospheric processes underlying the ultimate isotopic composition. A full appreciation of the meteorological underpinnings of this more complicated relationship remains elusive, but such knowledge will provide insight into how changes in high-latitude storminess may change during warmer climates. ■

David Noone is in the Department of Atmospheric and Oceanic Sciences, and the Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado 80309, USA.
e-mail: dcn@colorado.edu

1. Overpeck, J. T. *et al. Science* **311,** 1747–1750 (2006).
2. Sime, L. C., Wolff, E. W., Oliver, K. I. C. & Tindall, J. C. *Nature* **462,** 342–345 (2009).
3. Epstein, S., Sharp, R. P. & Goddard, I. *J. Geol.* **71,** 698–720 (1963).
4. Dansgaard, W. *et al. Science* **218,** 1273–1277 (1982).
5. Dansgaard, W. *Tellus* **16,** 436–468 (1964).
6. Jouzel, J. *et al. J. Geophys. Res.* **102,** 26471–26487 (1997).
7. Masson-Delmotte, V. *et al. J. Clim.* **21,** 3359–3387 (2008).
8. Worden, J. *et al. Nature* **445,** 528–532 (2007).
9. Noone, D. *J. Geophys. Res.* **113,** D04102 (2008).
10. Lee, J.-E., Fung, I., DePaolo, D. J. & Otto-Bliesner, B. *J. Geophys. Res.* **113,** D19109 (2008).

EPIGENOMICS

Methylation matters

Dirk Schübeler

Genome-wide maps of methylated cytosine bases at single-base-pair resolution in human cells reveal distinct differences between cell types. These maps provide a starting point to decode the function of this enigmatic mark.

Methylated cytosine, often referred to as DNA's fifth base, makes up a subset of nucleotides in the mammalian genome. As cytosine methylation does not affect base pairing, deposition of this mark can regulate processes, such as transcription, without affecting the genetic blueprint. Once established, the methylated cytosine modification can be faithfully copied to newly synthesized DNA, and so can be passed on to daughter cells, making it a true epigenetic mark. Many groups have studied the genomic distribution of DNA cytosine methylation and other chemical modifications of histone proteins to describe what has been dubbed the epigenome. On page 315 of this issue, a team headed by Joseph Ecker — Lister *et al.*[1] — provide the first complete DNA-methylation map of the human genome at single-base-pair resolution. Their accomplishment reveals intriguing features of the methylcytosine mark that were not identified in previous, less-comprehensive maps[2,3].

Many eukaryotes (plants and animals) use DNA cytosine methylation to silence parasitic elements that have invaded their genome, such as transposons and retroviruses. However, the mark can also be used to regulate the expression of naturally occurring genes, allowing an expanded repertoire of tissue-specific gene transcription. Because DNA cytosine methylation patterns change with age and in certain diseases (particularly in cancer), the distribution, function and regulation of this modification are of great interest as a diagnostic marker and potential therapeutic target.

To achieve such a high-resolution genomic map of DNA methylation, Lister *et al.*[1] treated genomic DNA from human cells with sodium bisulphite, which converts non-methylated cytosines to uracil, but leaves methylated cytosines untouched[4]. The authors then sequenced the whole genome multiple times — a tour de force that was made possible through the recent development of high-throughput sequencing technologies. In total, 178 gigabases of sequence were generated, the equivalent of sequencing the entire genome 57 times. Sequencing at this unprecedented scale was crucial to these studies, as it is inherently difficult to map bisulphite-converted sequences back to the genome. Furthermore, only multiple sequence reads allow for confidence in base allocation — here, an astounding 94% of all cytosines in the genome were identified.

Such sequence coverage not only provides an outstanding information resource, but detailed analysis of these data also reveals several surprises, particularly in the differences in cytosine methylation patterns between certain cell types. The authors chose the cells they compared wisely: a human stem cell, which is pluripotent and so can develop into any other cell type; and a fibroblast, which is fully differentiated. Although the observed methylation patterns are comparable between both cell types, there are remarkable

CHEMICAL PHYSICS

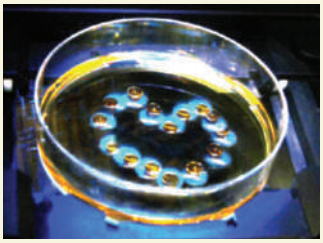
Guiding light

With microfluidic devices gaining prominence for many applications in chemistry and biology, the hunt is on to find ways of accurately controlling the motion of liquid droplets. In *Angewandte Chemie*, Antoine Diguet *et al*. describe a method for using light to trap and move oil droplets floating on an aqueous solution (A. Diguet *et al*. *Angew. Chem. Int. Edn* doi:10.1002/anie.200904868; 2009).

This isn't the first time that light has been used to push droplets around. But Diguet and colleagues take a new approach based on the chromocapillary effect, in which light generates a tension gradient at a liquid–liquid interface. This gradient can induce an interfacial flow between droplets and bulk liquids, which propels the droplet in the opposite direction to the gradient.

The authors' technique depends on the compound dissolved in the bulk liquid. Diguet *et al*. used a surfactant that isomerizes in response to different wavelengths of light — it adopts a polar isomeric form when illuminated with ultraviolet light, and a less polar form when lit with visible light. The light-induced changes in polarity modulate the surface tension between the surfactant solution and oil droplets floating on its surface. So, when the authors partially illuminated such a droplet with ultraviolet light, the tension gradient caused the droplet to move away from the lit area. If they then partially irradiated the droplet with visible light, the droplet moved towards the lit area.

By combining ultraviolet and visible light, Diguet *et al*. made a chromocapillary trap that captured oil droplets cast onto the surface of the surfactant solution. The authors could then drag the droplets across the surface of the solution, at speeds of about 300 micrometres per second, simply by moving the trap around. The image above is a montage of superimposed frames from a movie, and shows a droplet (gold colour) being directed by a trap (cyan halo) along a heart-shaped path; the Petri dish is 5.1 centimetres in diameter.



Chromocapillary traps should work for various combinations of immiscible liquids, and could thus be useful for controlling droplets in micro- or millifluidic devices. The authors' system could also be used to safely handle dangerous liquids, or in light-responsive materials.

Andrew Mitchinson

differences. Many genomic regions vary in the density of methylation of cytosines that lie immediately 5′ to a guanosine (known as CpG methylation), with the differentiated cells having large, prevalent, partially methylated tracts of DNA, which are associated with the reduced activity of genes that lie 5′ to these tracts. In fibroblasts, 99.98% of all methylation occurs at CpG dinucleotides (Fig. 1). This is not unexpected given that these dinucleotides are believed to be the exclusive target of DNA methylation in vertebrates.

Surprisingly, however, in the stem cells studied, around 25% of the methylation sites do not occur in the context of CpGs, but rather are found on cytosines that neighbour other bases, in particular adenosine (Fig. 1). Non-CpG methylation had been observed before[5] in mouse stem cells, but its prevalence in the genome was not widely appreciated and its genomic location was unclear. Lister *et al*.[1] show that its frequency varies between individual cells and is relatively low — only a small percentage of non-CpG cytosines in stem cells is methylated. Yet their genome-wide analysis also reveals that non-CpG cytosine methylation is enriched at active genes, specifically on the DNA strand that serves as a template for transcription. The authors speculate that this differential targeting is linked to the process of active transcription, reminiscent of the targeting of non-CpG methylation to expressed genes in the plant *Arabidopsis thaliana*[6,7]. Although the function and enzyme (or enzymes) responsible for non-CpG methylation are yet to be identified, this mark remains a curiously exclusive feature of stem cells. If terminally differentiated cells that lack non-CpG methylation are engineered to become induced pluripotent stem cells, they regain this unusual modification at the few loci tested by Lister and colleagues. Only further functional studies will reveal the specific role this mark has in stem-cell biology.

The work by Lister *et al*.[1] provides a milestone in the quantitative description of mammalian DNA cytosine methylation and highlights the dynamic nature of this mark during cell differentiation. The maps they have generated reveal that our understanding of the establishment and function of DNA methylation patterns is far from complete. Most notably, the question remains as to what extent the observed differences are consequences of differential gene activity or are actively involved in transcriptional regulation.

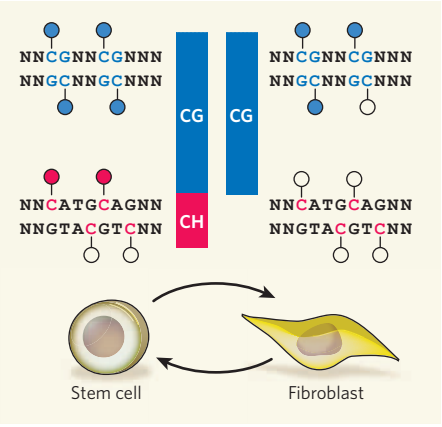


Figure 1 | DNA methylation patterns differ between stem cells and differentiated cells[1]. In stem cells, regions of DNA with CpG methylation (blue) are mostly uniformly methylated, whereas this modification is more heterogeneous in fibroblasts. Non-CpG methylation (red), which occurs primarily at CA nucleotides, is detected only in stem cells, yet is asymmetric and more scarce and patchy than CpG methylation. If fibroblasts are converted to induced pluripotent stem cells they regain non-CpG methylation. Filled circles, methylated cytosines; unfilled circles, unmethylated cytosines. H stands for A, C or T; N stands for any nucleotide.

The fact that DNA cytosine methylation patterns are cell-type specific and variable has led to the proposal that cytosine methylation may function as a memory module of cell identity and developmental state[8]. The feasibility of measuring complete DNA methylomes at the base-pair level provides the technical starting point to address this hypothesis in a quantitative and unbiased manner. Given the current cost of sequencing, these are still expensive experiments. Nevertheless, owing to the dynamic nature of DNA methylation, it is clear that we will appreciate the complexity of the distribution of this mark only after generating additional methylome maps from many distinct cell types from different individuals. Furthermore, unravelling the functional basis of DNA methylation will require combining such descriptive sequencing efforts with mechanistic studies. Global initiatives in defining genetic variations in humans provide a framework for how these endeavours can be achieved. Such coordination has already been initiated in the United States by the National Institutes of Health's Roadmap Epigenomics Program. And efforts are under way to coordinate an international initiative[9] to ultimately decode the function of this still-enigmatic base modification. ■

Dirk Schübeler is at the Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, 4058 Basel, Switzerland.
e-mail: dirk@fmi.ch

1. Lister, R. *et al. Nature* **462,** 315–322 (2009).
2. Meissner, A. *et al. Nature* **454,** 766–770 (2008).
3. Eckhardt, F. *et al. Nature Genet.* **38,** 1378–1385 (2006).
4. Frommer, M. *et al. Proc. Natl Acad. Sci. USA* **89,** 1827–1831 (1992).
5. Ramsahoye, B. H. *et al. Proc. Natl Acad. Sci. USA* **97,** 5237–5242 (2000).
6. Zhang, X. *et al. Cell* **126,** 1189–1201 (2006).
7. Zilberman, D. *et al. Nature Genet.* **39,** 61–69 (2007).
8. Mohn, F. & Schübeler, D. *Trends Genet.* **25,** 129–136 (2009).
9. Jones, P. A. *et al. Nature* **454,** 711–715 (2008).

Q&A

MALARIA

Evolution in vector control

Yannis Michalakis and François Renaud

Each week some 20,000 people die from malaria. There will be no magic ways of reducing this dreadful toll, not least because the mosquito vector and the parasite itself have formidable abilities to resist control measures. Angles of attack that rest on evolutionary principles are being explored.

Why hasn't malaria been eradicated?

The reason for the failure is threefold. First, there are no drugs that can cure people on a routine basis and on a large scale. Second, there are as yet no vaccines that can provide protection against the unicellular *Plasmodium* parasites that cause the disease. And finally, no large-scale, effective yet environmentally respectful way of controlling the vectors — the mosquitoes that transmit the parasites — has yet been applied in areas of intense malaria transmission. Political, social and economic factors also contribute to the burden imposed by malaria and to the failure to eradicate it.

What is the basic biology involved?

To complete its complex life cycle, *Plasmodium* needs to infect a vertebrate and then a mosquito host (Fig. 1). For human malaria, the vertebrate host is man, but other mammals, birds and reptiles can be infected by their own malaria-causing parasite. Only female mosquitoes transmit the disease. They become infected when they take a blood meal at night, and pass the parasite on during a meal roughly 10–14 days later; the disease symptoms of chills and fever are caused by *Plasmodium* proliferation in red blood cells. More than 20 species of the mosquito genus *Anopheles* can transmit human malaria, although with variable efficiency. A notable feature is that even in areas where most, if not all, humans are infected, only a small proportion of mosquitoes are.

What are the ways of trying to combat malaria?

There are two broad approaches: direct, involving vaccination to prevent infection or drugs to cure it; and indirect, involving attempts to reduce or stop transmission by the mosquito vector. Vector control is the subject of this article.

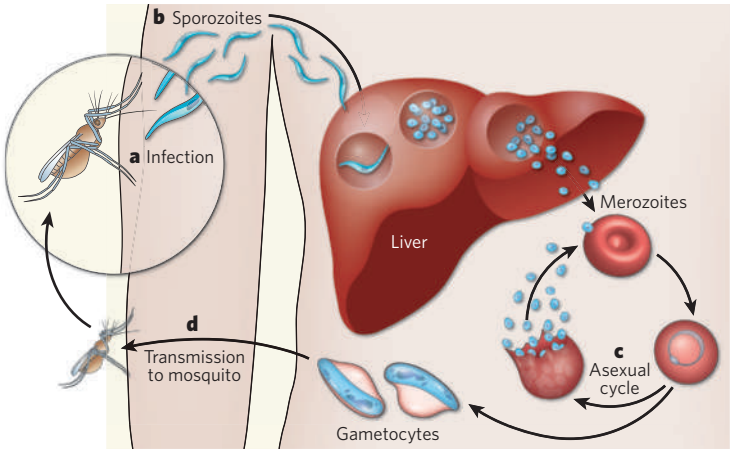


Figure 1 | Basic features of the *Plasmodium* life cycle. **a**, Egg development in female *Anopheles* mosquitoes requires a blood meal. In the process, infected females inject the sporozoite form of the parasite into a human host. **b**, Sporozoites are carried in the bloodstream to liver cells, where they proliferate asexually, and then, as merozoites, invade red blood cells. **c**, An asexual cycle within red blood cells, at which disease is clinically manifest as fever and chills, is followed by the production of male and female gametocytes. **d**, These are transmitted back to a mosquito during a blood meal, where they fuse to form oocysts that duly divide to create sporozoites. These migrate to the salivary glands, where the cycle of infection starts again. Details of the life cycle differ for different species of *Plasmodium*, with *P. falciparum* causing the most virulent form of human malaria. Development within the mosquito is temperature-sensitive, and takes 10–14 days or longer.

Why emphasize vector control?

The development of drug resistance within populations of *Plasmodium* and the inability, so far, to formulate effective vaccines have limited the direct ways of tackling malaria. Another reason for revisiting vector control is that many people carry and transmit the parasite, but don't show the symptoms of malaria and are not treated. Finally, vector control is historically the cheapest and most successful approach, and evolutionary biology is now offering fresh perspectives on it.

What are the approaches to vector control?

They aim either at lowering vector density or at avoiding infectious bites. The first large-scale programmes were attempts to destroy the habitat of mosquito larvae by draining wetlands. But this approach is costly, and mosquitoes can breed in small bodies of water that cannot be eliminated. Killing larvae or adults with chemical insecticides is another approach, and there have also been studies with biological-control agents such as nematodes and fungi. With the development of molecular biology came the hope that the release of genetically manipulated mosquitoes would prove effective.

How are insecticides used?

Spraying with chemical insecticides has been the commonest form of mosquito control because it has proved effective, at least locally. Application of DDT, for example, causes significant decreases in mosquito densities. Other products, both artificial and natural (such as bacterial toxins), directed towards larval or adult mosquitoes, are widely used today. Insecticide-impregnated bednets have a double advantage: they both protect people from infectious bites and reduce vector populations.

What are the problems with insecticides?

Insecticides are typically toxic not only to vectors but to many other organisms, in some cases including humans. Moreover, sooner or later mosquitoes develop resistance to insecticides, either by becoming able to metabolize them, or by modifying the insecticide's target molecule. Typically, resistance comes at a fitness cost to mosquitoes, but the vectors have been able to counteract pretty much any insecticide used against them. In evolutionary terms, the benefit from resisting insecticides

at lethal doses is always much larger than the accompanying cost. Surprisingly, we do not know whether insecticide-resistant mosquitoes are more or less effective malaria vectors.

How does genetic modification enter the picture?

The rise of insecticide resistance prompted research into the release of genetically manipulated mosquitoes (GMMs), with the aims of limiting *Plasmodium* transmission or reducing mosquito population densities. One strategy is best known from attempts to release male mosquitoes that can mate but cannot reproduce successfully, although other approaches, such as the use of site-specific 'selfish genes', have been proposed (Box 1). Another involves producing mosquitoes in which *Plasmodium* development is halted at any stage.

What are the problems with GMMs?

Apart from the political controversy associated with any use of genetically modified organisms, there are practical drawbacks. The success rate of genetic-transformation techniques is low, leading to possible inbreeding-related decreases in GMM population viability. The genetic manipulations often involve laboratory strains, whose ability to survive in the wild, let alone invade wild mosquito populations, is questionable. It is not clear that GMMs designed to resist *Plasmodium* infection would necessarily have a fitness advantage, or, given the low prevalence of mosquito infection even in areas where many humans are affected, that such resistance would spread rapidly. For most, though not all, proposed GMMs, their spread and maintenance in the wild have to be ensured by the repeated release of modified mosquitoes; strategies based on naturally invasive driving mechanisms, such as homing endonuclease genes or transposable elements, present an advantage in this respect (Box 1). There is also the complication that malaria is transmitted by many mosquito species, and GMMs might have to be designed for each of them or restricted to the main culprits. Moreover, if genotype–genotype (mosquito–*Plasmodium*) interactions exist at the within-species level, modification of a given gene might affect *Plasmodium* only on a local scale. Finally, release of *Plasmodium*-resistant mosquitoes might increase transmission of other *Anopheles*-borne diseases, such as filariasis.

How about enlisting the natural enemies of mosquitoes?

Such organisms include nematodes that kill mosquito larvae, microsporidia that infect larvae, and fungi that attack adults (Box 2, overleaf). The fungal parasites can induce more than 80% mosquito mortality within 14 days of infection (a critical period, as we will see later), and are especially effective against *Plasmodium*-infected mosquitoes. They can be sprayed indoors, on places where mosquitoes rest before or after blood meals, and can be combined with other control or prevention strategies, such as bednets.

Box 1 | Approaches involving genetically modified mosquitoes (GMMs)

The production of sterile male mosquitoes can be achieved non-genetically through radiation treatment, or by introducing dominant lethal genes into them. This works, for example, by introducing genes that are lethal during larval development, unless the mosquitoes are fed a specific substance such as an antibiotic. The GMMs survive in the lab because they are fed the antibiotic. But when they are released and then mate with wild females, their offspring die because of the absence of the antibiotic in the environment.

An approach that involves 'selfish genes' entails, for example, the use of homing endonuclease genes (HEGs) that encode an enzyme that recognizes a 20–30-base-pair sequence on chromosomes not containing the HEG, and cleaves it. The cleaved copy uses the HEG-carrying copy as a repair template, and thus the HEG spreads in the population. Because the HEG lies in the middle of the recognition sequence, the chromosome carrying it is protected from future attack.

For use in vector control, HEGs are required that recognize and insert into a specific sequence of an essential mosquito gene. When these GMMs are heterozygotes, with one of a gene pair disrupted and the other not, they will survive and reproduce. Such constructs will initially spread because they are driven by the bias in HEG transmission. Once their frequency is sufficiently high in the mosquito populations, substantial numbers of inviable homozygotes, in which both copies of the gene are disrupted, will result, possibly causing the vector population to crash.

GMMs are also used to spread *Plasmodium* resistance among mosquitoes. The first stage consists of identifying genes that stop *Plasmodium* development in the mosquito, or prevent transmission, and then engineering mosquitoes that express this characteristic. This is the relatively easy part. The problem then consists of associating this construct with a naturally invasive driving mechanism, which could involve transposable elements, meiotic drive, HEGs or bacteria called *Wolbachia*.

In principle, this approach does not harm the vector, only *Plasmodium*, and the driving mechanisms may help to overcome the other disadvantages related to the release of GMMs. But the association between the resistance gene and its driver may be broken by genetic recombination, and some of the drivers — such as transposable elements — may generate insects with undesirable characteristics. Also, it is difficult to assess the risk, and the consequences, of the driving mechanisms and transgenes spreading to other species. **Y.M. & F.R.**

... and the drawbacks are?

Like GMMs, mosquito predators and parasites must spread and be maintained in wild populations, while not eliciting resistance in the target. The spread has to be achieved by human agency, and so requires the production of large quantities of these agents; the existence of industrial formulations for some of them, for example fungi or nematodes, is a definite plus. If rearing these natural allies is relatively easy, it can also satisfy the demand for maintenance, although any ally able to meet that demand itself would obviously be a favourable option. The development of resistance is much more of a problem. Widespread use of natural allies would drastically shift selection in favour of mosquitoes able to resist them, as has happened with chemical insecticides. The consequences, including the ensuing selective pressure imposed on *Plasmodium*, can neither be generalized nor ignored. A further consideration is the specificity, or lack of it, of natural allies. Perhaps killing any insect present in a human habitation is a good thing, but this too is an issue that merits debate. Alternatively, mosquito-specific isolates could be selected and used.

How does the timing of intervention come into things?

Like some other vector-borne agents, such as the virus that causes dengue fever, *Plasmodium* has an intriguing feature: after it has entered a mosquito, it takes quite a long time, 10–14 days or more, to produce the stages transmissible to humans (Fig. 1). The timing of the effect of an insecticide, whatever its nature, may thus be crucial. An agent that acts once transmission has taken place will not be very helpful; and one that imposes a large burden on mosquito fitness will favour any mechanism that makes mosquitoes resistant. But get the timing of intervention right, and it might be possible to dissociate the effects on *Plasmodium* transmission from the effects on mosquito fitness.

How might this influence anti-vector strategies?

Until recently, all such strategies worked on the assumption that negative effects on vector fitness are desirable. But what we really want to do is limit pathogen transmission. An ideal approach would cause as little harm as possible to the vector, to avoid eliciting resistance, but as much harm as possible to the pathogen. This is where the evolutionary theory of senescence comes in.

What is the evolutionary theory of senescence?

The strength of natural selection declines with age. In consequence, selection against mutations with negative effects late in life is much weaker than selection against mutations with negative effects early in life. From an evolutionary standpoint, senescence can thus be explained either

by the accumulation of mutations having deleterious effects late in life (the 'mutation accumulation' theory; Fig. 2), or by the fixation of mutations that are advantageous to their bearer when it is young and detrimental when it is old (the 'antagonistic pleiotropy' theory).

How can this theory be applied in vector control?

A consequence of the fact that *Plasmodium* requires a long developmental period within mosquitoes is that malaria is transmitted only by relatively old mosquitoes. Females undergo cycles during which a blood meal is necessary to produce eggs, which are laid on water, with each cycle lasting 2–4 days. Because their daily survival rate is 80–90%, most females will go through few such cycles before they die (fewer than 20–40% would go through more than four cycles). Thus, a strategy killing mosquitoes later in their life, but before they transmit malaria, would mimic senescence and disrupt transmission. Such approaches would generate little, if any, selection for resistance in the mosquito population, and would require 'late-life-acting insecticides'.

Can this theory be put into practice?

Pathogenic fungi seem to have the desirable properties: they induce high mortality relatively late in a mosquito's life but before the insect transmits malaria (Fig. 2). These fungi are even more virulent in malaria-infected mosquitoes, which is a bonus: this should further slow the advent of resistance to the fungi while at the same time favouring resistance to *Plasmodium* among mosquitoes. The same principles are behind a project to combat dengue fever, a viral disease transmitted by *Aedes* mosquitoes, by introducing *Wolbachia* bacteria into the vector through genetic manipulations that have little effect on vector fitness but that can disrupt dengue transmission. The manipulation of malaria-transmitting mosquitoes in like fashion is under way. *Wolbachia* could in principle spread in mosquito populations because there is a driving mechanism: infected males effectively sterilize females uninfected by the same bacteria as the males. But although this strategy is based on the same evolutionary principles as that involving the fungi, there are two drawbacks. First, large-scale application in the field will encounter the same issues as other approaches involving GMMs. Second, there are examples from *Wolbachia* interactions with fruitflies, butterflies and even mosquitoes that show that bacterial virulence and/or within-host density evolves rapidly. So frequent release of mosquitoes carrying 'unevolved' bacteria could be required.

How might *Plasmodium* respond to fungal attack on its vector?

One way would be for the parasite to shorten its development time within the mosquito. Here we enter some fascinating and largely uncharted territory. We can assume that the present long development time is advantageous, but if late-life-acting insecticides were to be widely applied, the selection 'landscape' for *Plasmodium* would change dramatically. Genes that determine slower development would become a liability, and this prompts various questions. Is there a link between *Plasmodium* development time and virulence in mosquitoes? If faster development leads to lower virulence, there would be selection against any naturally occurring resistance mechanisms in the mosquitoes, and parasite transmission would be enhanced. And is there a link between development time in mosquitoes and infectiousness and/or virulence in humans? Selection experiments on varying *Plasmodium* development times should allow us to explore these questions.

What's next?

By dissociating fitness effects on the vector from effects on parasite transmission, late-life-acting insecticides offer a new angle of attack on malaria by limiting the potential evolution of insecticide resistance. The approach will require more testing of both theory and practice, and proven methods, such as insecticide-treated bednets or insecticide spraying, will remain the mainstays for malaria control. Although much research is devoted to the separate development of control strategies, that is perhaps at the expense of assessing the combinations of approaches that could produce the best results in designing evolution-proof inhibition of *Plasmodium* transmission. Papers published in the past couple of months show that a combination of conventional insecticides and insect-killing fungi can work synergistically in directly lowering malaria transmission and decreasing insecticide resistance. Evolutionary thinking can greatly contribute to disease control, particularly in devising ways to limit disease transmission, to avoid the development of resistance and to predict the potential evolutionary responses of parasites and vectors. ■

Yannis Michalakis and François Renaud are in the Laboratoire de Génétique et Evolution des Maladies Infectieuses, UMR CNRS IRD 2724, IRD, 34394 Montpellier Cedex 5, France.
e-mails: yannis.michalakis@mpl.ird.fr; francois.renaud@mpl.ird.fr

Box 2 | Fungal allies in vector control

The fungi concerned are types of Sordariomycetes, and the species of principal interest are *Beauveria bassiana* and *Metarhizium anisopliae*. The fungal spores attach to the insect cuticle when, for instance, a mosquito rests on a wall that has been sprayed with an appropriate preparation. The spores then germinate and penetrate the cuticle, after which they develop in the insect's haemocoel (the equivalent of the circulatory system). The fungi breach the cuticle using enzymes, and then overcome the insect's immune system either by producing cryptic forms that the immune system does not 'see', or by secreting substances that suppress immunity. Ultimately, the fungus kills its host and produces spores.

It is unlikely that the fungi will be able to self-sustain and provide long-term infectivity; repeated applications will probably be necessary. Spore viability, a critical component for the success of this approach, depends on ambient temperature and humidity, and variation among fungal strains for spore viability is poorly understood. The effects of these fungi on mosquitoes have been investigated in the lab and in a pilot study in Tanzania, and formulations are being developed for use in the field. Such work may profit from previous experience, because the same fungi are used as large-scale biological-control agents against locusts and grasshoppers. **Y.M. & F.R.**

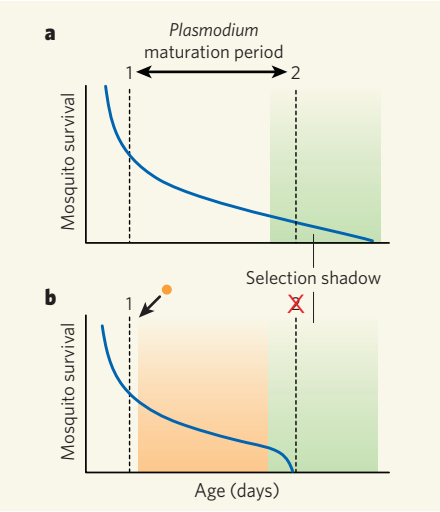


Figure 2 | The mutation-accumulation evolutionary theory of senescence and its application in malaria control. a, Survival of an adult female mosquito that was infected by *Plasmodium* during her first blood meal (1) and transmitted *Plasmodium* during a subsequent blood meal (2). The interval between the two meals corresponds to the *Plasmodium* maturation period, during which the female may have taken other blood meals without transmitting the parasite. For simplicity, we assume fecundity does not vary with age. Because survival declines with age, selection is much weaker against mutations with deleterious effects late in life — shown by the 'selection shadow'. (Modified after T. B. L. Kirkwood & S. N. Austad *Nature* 408, 233–235; 2000.) **b**, Survival of a female mosquito infected by *Plasmodium* and a fungus (orange dot) at her first blood meal. The fungus grows in the female (orange shaded area), which dies before she can transmit *Plasmodium*. In effect, the fungus acts as a late-life-expressed deleterious mutation.

FURTHER READING
www.mosquitoage.org
www.nature.com/nature/supplements/collections/malaria
www.thereadgroup.net
www.malariaworld.org
Rose, M. R. *Evolutionary Biology of Aging* (Oxford Univ. Press, 1995).
www.rollbackmalaria.org
go.nature.com/PiiHxJ

REVIEWS

Light and shadow from distant worlds

Drake Deming[1] & Sara Seager[2]

Exoplanets are distant worlds that orbit stars other than our Sun. More than 370 such planets are known, and a growing fraction of them are discovered because they transit their star as seen from Earth. The special transit geometry enables us to measure masses and radii for dozens of planets, and we have identified gases in the atmospheres of several giant ones. Within the next decade, we expect to find and study a 'habitable' rocky planet transiting a cool red dwarf star close to our Sun. Eventually, we will be able to image the light from an Earth-like world orbiting a nearby solar-type star.

The first exoplanet found to orbit a solar-type star (51 Peg) was a startling discovery[1]. This gas-giant planet of half a Jupiter mass orbits its star at a distance six times closer than the radius of Mercury's orbit in our own Solar System. The exoplanet, 51 Peg b, was discovered by measuring the line-of-sight (radial) velocity of the star as it orbited the centre-of-mass of the system. The magnitude of this velocity reflex yields—via conservation of momentum—the planet mass times the sine of the orbital inclination. Given an astronomical estimate of the stellar mass, the semi-major axis of the planet's orbit follows from Kepler's law. At an orbital distance of 0.05 astronomical units (1 AU is the Earth–Sun distance), 51 Peg b should be heated by stellar irradiation to a temperature in excess of 1,000 K. The astonishing close orbits and incredibly high planetary temperatures for the so-called hot Jupiters shattered the Solar System paradigm of planet formation and was the first surprising discovery of many in exoplanetary science.

The most exciting goal in exoplanetary science is to find and characterize a rocky habitable planet like our Earth. Although we have not yet found a planet that matches Earth in terms of habitability, we have measured light from exoplanets—especially the hot Jupiters—and we are beginning to characterize the properties of their atmospheres. This review concentrates on the bulk properties (masses, radii) and atmospheric compositions of exoplanets, with an eye towards the eventual characterization of a rocky, habitable world.

The best summary of exoplanets is their distribution of mass versus the semi-major axis of their orbit (Fig. 1). Most significantly, exoplanets span all ranges of mass and semi-major axis available to current detection techniques. This is a consequence of both the stochastic nature of planet formation, and of planet migration through the protoplanetary disk. Regions of Fig. 1 are blank because of selection effects and technological limitations. Fewer planets detected by radial velocity are found at large orbital distances (>5 AU) in Fig. 1 because most of these planets have not yet completed a sufficient portion of their >12-year orbits for the detection to be finalized. Also, the sensitivity of the radial velocity surveys is currently limited by stellar activity to about 1 m s^{-1} in velocity amplitude. This precludes the detection of Earth-twins at 1 AU, but allows the detection of rocky planets as small as several Earth masses orbiting close to their host stars. These so-called super-Earths are loosely defined to be rocky or icy planets between 1 and 10 Earth masses, and can have radii twice that of our Earth, or more. Current research aims to detect and characterize a habitable super-Earth orbiting a nearby low-mass star within the next decade. Over a longer term, the detection and characterization of exoplanets that are very similar to our Earth may be possible, using advanced imaging techniques.

Direct detection and characterization of exoplanets

In spite of the astounding success of the radial velocity technique[2], this method only measures the wobble of the star and does not detect planets directly. In other words, the radial velocity technique does not measure light from the planets. To characterize the nature of the planet's atmosphere, we must invoke techniques that isolate the light emitted from, or reflected by, the planet. An extension of traditional astronomical techniques to observe a planet spatially resolved from the star is high-contrast imaging, as being developed for a Terrestrial Planet Finder mission[3,4]. Recent advances in ground-based imaging[5,6] have led to the discovery of giant planets orbiting dozens to hundreds of astronomical units from young massive stars—shown in the upper

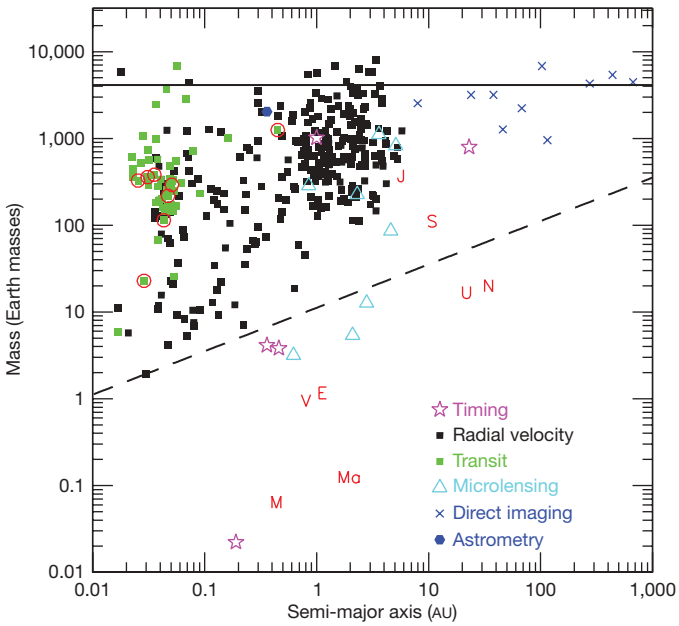


Figure 1 | Distribution of known exoplanets in mass and orbital semi-major axis. Planets detected by different techniques are shown using different symbols. The symbols circled in red are those planets with an analysis of their atmosphere, published as of September 2009. The solid horizontal line is the nominal upper mass limit above which an object is not considered to be a planet. The dashed line represents a radial velocity reflex of 1 m s^{-1} for a solar-type star, and planets orbiting solar-mass stars producing smaller velocity signals are generally not detectable using the radial velocity technique. Red letters indicate Solar System planets: M, Mercury; V, Venus; E, Earth; Ma, Mars; J, Jupiter; S, Saturn; U, Uranus; N, Neptune.

[1]Planetary Systems Laboratory, Code 693, NASA's Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. [2]Department of Earth, Atmospheric, and Planetary Sciences, and Department of Physics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02138, USA.

right of Fig. 1. The masses of planets at such large distances cannot be measured using radial velocities, so they are estimated by comparison of the planet brightness to cooling curves for young planets[7], or by their gravitational effect on disk structure[8].

A major challenge in imaging exoplanets is that the ratio of their brightness to that of their star is very low (from 10^{-3} to 10^{-10}, depending on planet and stellar types, and wavelength). Extension of large-separation, giant planet imaging to Earth-like planets must bridge several orders of magnitude in planet–star brightness and a factor of 10 or more in planet–star orbital separation (Fig. 1), and will require space-borne measurements. Although dramatic progress has been made in the laboratory[3], a Terrestrial Planet Finder mission is a decade or more in the future. Meanwhile, significant progress is being made in direct detection and characterization of exoplanets using the transit technique.

Shadows of distant worlds

The nearness of the hot Jupiters to their stars means that they have a significant probability (typically about 0.1) of transiting their star as seen from Earth. During the transit, our Earth falls within the shadow of the exoplanet, and the light we receive from the parent star is diminished by a small amount. The occurrence of a transit—when the planet passes in front of its star as seen from Earth—is a great advantage to physical characterization of the planet[9], as illustrated in Fig. 2. High precision photometry during transit can be used to measure the blocking of stellar light versus time, and this so-called transit light curve is sufficient to determine the radii of both the planet and star, if the stellar mass can be estimated (for example, by using stellar models and the star's colours). Moreover, the planetary and stellar radii are proportional to the cube-root of the stellar mass, and are thus minimally sensitive to errors in the adopted stellar mass value. Nearly every transiting planet host star has also been measured by radial velocity, so both the planet's radius and mass are known.

The first transiting exoplanet[10,11] revealed a mystery that has so far defied explanation. The observed mass–radius diagram for transiting giant planets (Fig. 3) shows that several hot Jupiters have radii significantly greater than predicted[12]. Although a reduced size of the heavy element core in a giant planet allows a larger radius, this is not enough to explain the inflated radii of some giant planets, as shown in Fig. 3. Some process must be at work that generates energy in the interiors of these planets, inflating their radii, and possibly perturbing other aspects of their interior structure, in ways that we do not understand. One possibility that is currently the subject of debate is

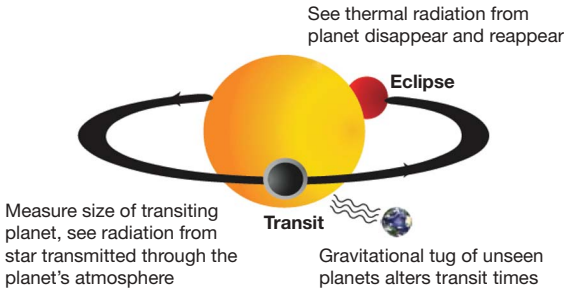


Figure 2 | Geometry and science yield from transiting planets. During transit, the fraction of stellar light blocked by the planet (shown black), together with the detailed shape of the transit curve, yields the radius of both the star and the planet. Stellar light transmitted through the annulus of the atmosphere (grey ring around black planet) reveals atomic and molecular absorption features from the planet's atmosphere. During eclipse, the disappearance of the planet (shown red) allows the stellar light to be measured in isolation, and subtracted from the total light of the system to yield the light from the planet, and its spectrum. Note also that the presence of unseen planets can in principle be inferred from variations in transit times (and also eclipse times). Representative Earth at bottom right indicates possible unseen planets that could perturb transit times.

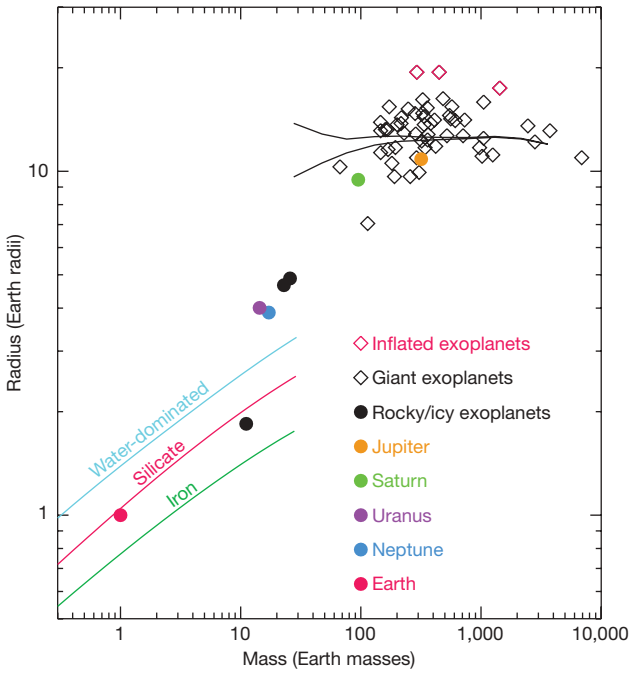


Figure 3 | Mass–radius diagram for transiting planets. Black diamonds are giant exoplanets. Three inflated giant exoplanets are indicated as red diamonds; from left to right they are: TrES-4, WASP-12 and OGLE-TR-L9. The lower black line is a theoretical mass–radius relation[74] for 1-Gyr-old giant planets orbiting at 0.045 AU from a solar-type star, and having a 10-Earth-mass core of heavy elements, plus a hydrogen–helium envelope. The upper black line is the same, but with zero core mass. The rocky/icy exoplanets (black filled circles) are the two exo-Neptunes (GJ 436b and HAT-12), and the super-Earth CoRoT-7b[22]. The coloured lines are theoretical mass–radius relations[66] for super-Earths lacking a hydrogen–helium envelope, but having a solid composition that is either water-dominated, silicates, or iron. Note the positions of Solar System planets, including Earth.

remnant heat from cosmically recent tidal circularization of their orbits[13]. Timing the eclipse of the planet (see below) shows that the orbits are often very close to circular[14,15], suggesting that the tidal evolution of the orbital shape is complete, but that may not preclude remnant internal energy from tidal dissipation. Another possibility (among many) is that convective transport of energy from the planetary interior may be less efficient[16] than calculated in simple homogenous models, owing to inhomogeneities in mean molecular weight.

Detection of transits of HD 209458b galvanized the astronomical community to find more transiting planets, and currently more than 45 planets are known to transit stars brighter than thirteenth visual magnitude. Most of these transits were discovered by photometric surveys, and the discovery rate has exploded in the past few years[17–19] as the transit technique matured and groups learned to cull their candidates and eliminate false-positives. Transit surveys have now discovered two planets comparable to Neptune in size[20,21], and one super-Earth only 70% larger in radius than our own Earth[22]. The recent launch of NASA's Kepler mission[23] will greatly increase the number of rocky and/or icy transiting planets known. Most exciting is that Kepler is designed to tell us the frequency of true Earth analogues—by detecting Earth-sized planets orbiting in the habitable zones of Sun-like stars.

Light from distant worlds

Planets that transit their star will also pass behind the star (for planets on circular orbits). This eclipse of the planet by the star will occur approximately half an orbital period after transit (Fig. 2). Radiation from the planet can be measured from the modulation of the combined planet and starlight, because the planet's light is blocked out during eclipse and then later reappears. Such eclipse measurements

can be made with existing or planned facilities, because observatories optimized for general astronomy turn out to be suitable for transiting planet observations (for example, the Hubble Space Telescope and Spitzer Space Telescope). Hot Jupiters are expected to have temperatures between 1,000 and 2,000 K, so their atmospheres emit infrared radiation and the emitted flux peaks in the relatively near infrared wavelength region (1–5 μm). Successful detection of infrared light from exoplanets was first accomplished via eclipses observed using the Spitzer Space Telescope[14,24]. Spitzer has observed eclipses for more than two dozen exoplanets, and one of the best-observed[25–27] examples (HD 189733b) is shown in Fig. 4. Many additional in-depth analyses are becoming available at a rapid pace[15,28–31]. The Spitzer results, in combination with other key observations, have clarified several aspects of hot Jupiter atmospheres (but have also raised new questions) in three key areas as follows:

Hot Jupiters are both hot and dark. Hot Jupiters are blasted with radiation from the host star. The hot Jupiters should therefore be kinetically hot, heated externally by the stellar irradiance. Indeed, early hot Jupiter model atmospheres already predicted temperatures exceeding 1,000 K (refs 32, 33). The first and most basic conclusion from the Spitzer detections was the confirmation of this prediction[34]. The fact that the planets emit generously in the infrared implies that they efficiently absorb visible light from their stars. Searches for the reflected component of their energy budget have indicated that the planets must be very dark in visible light, with geometric albedos less than about 0.2 (refs 35, 36) and probably much lower. Purely gaseous atmospheres lacking reflective clouds can be very dark[33,37,38] but HD209458b also requires a high-altitude absorbing layer (see below) to account for its atmospheric temperature structure.

Water vapour in emergent spectra of hot Jupiters. A planetary atmosphere with elemental composition close to solar and heated upwards of 1,000 K is expected to be dominated by the molecules H_2, H_2O, and, depending on the temperature and metallicity, CO and/or methane (CH_4). Of these molecules, H_2O is by far the most spectroscopically active gas. Water vapour is therefore expected to be the most significant spectral feature in a hot Jupiter atmosphere. Some initial indications from Spitzer spectroscopy that water absorption was absent[39,40] were superseded by additional work that clearly showed water absorption[41–43]. Although models correctly predicted that the spectra of hot Jupiters are shaped by water absorption, the variation in temperature structure from one planet to another is not fully understood.

Other atoms and molecules identified in hot Jupiter atmospheres are atomic hydrogen[44], atomic sodium[45], methane[42] and carbon dioxide[43]. Atomic hydrogen is slowly escaping from HD 209458b, possibly forming a comet-like coma around the planet[44].

Day–night temperatures and thermal inversions. Hot Jupiters are fascinating fluid dynamics laboratories because they probably have a permanent dayside and a permanent nightside. Close-in giant planets are theorized to have their rotation synchronized with their orbital motion by tidal forces. Under this tidal-locking condition, the planet will keep one hemisphere perpetually pointed towards the star, with the opposite hemisphere perpetually in darkness. In the absence of atmospheric circulation, the star-facing hemisphere would be strongly heated, and the opposite hemisphere would be cold. There is evidence that such planets exist[46,47]. In contrast, other hot Jupiter exoplanets show relatively little temperature variation from the dayside to the nightside[27,47,48], and one particularly relevant Spitzer measurement is illustrated in Fig. 5. It seems likely that at least some hot Jupiters transport energy horizontally via zonal winds having speeds comparable to the speed of sound[49].

Figure 5 shows the Spitzer observations[27] of HD 189733b, where the strong winds have advected the hottest region to the east of the sub-stellar point. Giant planets in our Solar System also have strong zonal winds, appearing in multiple bands at different latitudes.

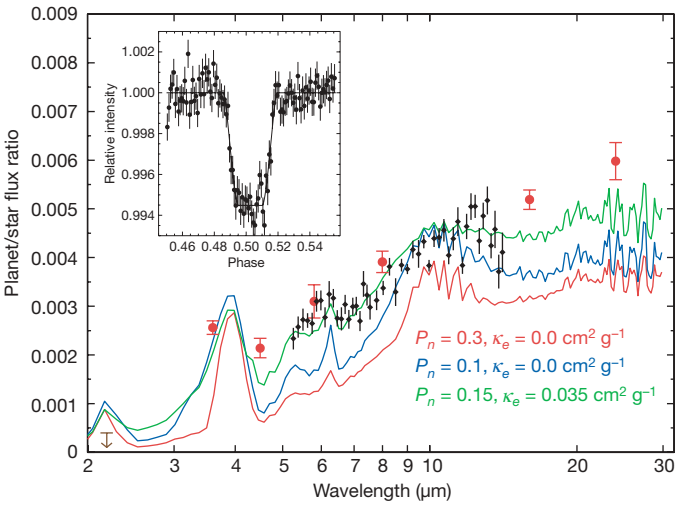


Figure 4 | Spitzer observations of the giant exoplanet HD 189733b. Observations of the planet are plotted as planet flux divided by stellar flux (stellar flux is reliably known). The red points with error bars (s.e.m.) are the results from eclipse depths measured[26] using photometry; the black points with error bars (s.e.m.) are from spectroscopy during eclipse[41]. The models (for example, ref. 75) have different amounts of assumed heat redistribution by winds (P_n parameter) and different opacities (κ_e) for a high-altitude absorbing layer. An example[25] of Spitzer eclipse photometry at 16-μm wavelength is shown in the inset. (Eclipses like that shown in the inset, but at multiple wavelengths, are used via photometry and spectroscopy to define the points that are compared to the models.) Main figure reproduced from ref. 41 with permission.

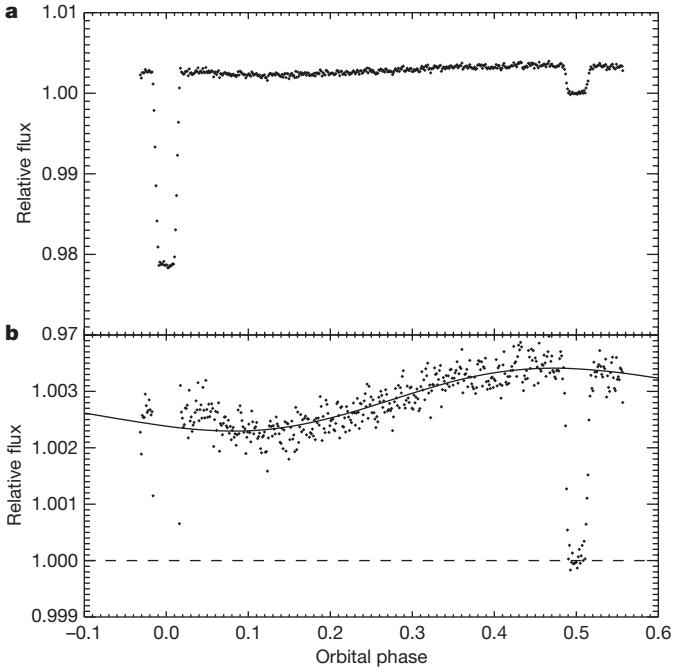


Figure 5 | Spitzer 'around the orbit' observations of the transiting exoplanet HD 189733b. **a**, The transit (at left) and the eclipse (at right) of the planet. **b**, Expanded intensity scale for the same observations as **a**, showing the variation of flux with orbital phase. Following transit of a planet having a tidally locked rotation, the planet's dayside will increasingly contribute to the combined light, resulting in greater flux. Note that the flux increase (dayside versus nightside) is only a fraction of the eclipse depth (dayside versus planet in eclipse), hence the flux difference between dayside and nightside is only a fraction of the planet's total infrared emission at this wavelength (8 μm). Note also that the peak flux from the planet occurs before eclipse, and that implies that the hottest region of the planet has been advected eastward by the strong zonal circulation. Figure reproduced from ref. 27 with permission.

Consideration of the relatively slow rotation rate of hot Jupiters (probably equal to their orbital period of a few days) leads us to believe that their zonal winds will occur predominately in one or two major jets that are quite extended in latitude and longitude[50]. The relatively large spatial scale of hot Jupiter winds—and the corresponding temperature field—should be a boon to their observational characterization. Planets with temperature fluctuations only on small scales will have little to no variation in the amount of their hemisphere-averaged flux as a function of orbital phase.

Many Solar System planets have thermal inversions high in their atmospheres (that is, the temperature is increasing with height above the surface). These so-called stratospheres are due to absorption of ultraviolet solar radiation by CH_4-induced hazes or O_3. Thermal inversions in hot Jupiter atmospheres were not widely predicted, because of the expected absence of CH_4, hydrocarbon hazes, and O_3. But Spitzer data show that the upper atmospheres of several planets do have thermal inversions[15,29–31], probably fuelled by absorption of stellar irradiance in a high-altitude absorbing layer. Possibilities for a high-altitude absorber include gaseous titanium dioxide and vanadium dioxide[51,52], as well as possibilities involving photochemical hazes[53]. Under a simple irradiation-driven scenario, the stronger the stellar irradiance, the more likely that an inversion would occur[51,52]. In addition, under such a scenario, planets with strong thermal inversions are also expected to show strong day–night temperature gradients. Hot Jupiters are probably more complex than this simple division allows; HD 189733b and XO-1b have virtually identical levels of irradiation and yet XO-1b has an inversion[30] while HD 189733b does not[26]. So it is possible that not-yet-understood chemistry may be a more dominant factor than stellar irradiance.

The growing diversity of exoplanets

The diversity of planet types available for study is expanding rapidly—planets with astonishing orbital and physical characteristics are announced with increasing frequency. Several close-in giant exoplanets have quite eccentric orbits, the eccentricity reaching to $e = 0.93$ in the case of HD 80606b. The stellar irradiance level received by a planet in an eccentric orbit can vary strongly with time. For example, HD 80606b receives a blast of radiation near periastron equal to 10,000 times the flux that our Earth receives from the Sun, and the planet is rapidly heated in response[54]. The blast at periastron lasts about two Earth-days, a small time frame compared to the 111-day orbit. We can learn much about the physics of the planet's atmosphere—and even its interior—by studying the atmosphere's response to such strongly varying stellar irradiance.

Much of the current excitement lies in the discovery and characterization of exoplanets different from the hot Jupiters that have been the main focus for the past decade. Core-dominated planets have interiors composed largely of iron, rock, or icy materials, which are fundamentally different to the predominant hydrogen/helium composition of hot Jupiters and other giant exoplanets. We anticipate a large sample of transiting core-dominated planets with measured masses and radii that will shed light on the nature and formation of rocky planets like our Earth. Masses and radii give average densities; together these can constrain the composition of planet interiors.

We are also beginning to study planets on more distant orbits, with correspondingly cooler atmospheres, and that are orbiting different kinds of stars (for example, stars of much lower and much greater mass than the Sun). One of the most fascinating recent exoplanet discoveries is the finding that some hot Jupiters appear to have retrograde orbits—the planets are revolving around the star in a different direction from that in which the star is rotating[55]. This is especially interesting, because planets form out of the protoplanetary gas and dust disk that is believed to rotate in the same direction as the star. These 'backward-orbiting planets' may have experienced a violent close encounter with another planet in the same system. Alternatively, the backward planet's orbit could have been gradually 'flipped' during planet migration by a more distant planet in the system[56,57].

New observing windows and techniques

Spitzer infrared observations have been the dominant source of characterization data for exoplanets to date, but new observational windows and techniques are being developed. Visible light thermal emission is predicted to be of significance for the hottest planets[58]. Eclipses of CoRoT planets[59] have been detected using visible light, and the Kepler team recently reported a spectacular measurement[60] of the eclipse of HAT-P-7b. Additional detections are expected from Kepler, and possibly from EPOXI[61]. Visible wavelength observations combined with infrared data help us to understand the relative contribution of thermal emission and reflected light, and will enable constraints on the planetary albedo. New ground-based photometric detections[62,63] of hot Jupiter eclipses are opening up a new wavelength range (2–5 μm). Combined with Warm Spitzer (that is, the non-cryogenic phase of the mission) observations and models, the new ground-based observations could enable constraints on the abundances of spectroscopically active gases and the presence and reflectivity of clouds or hazes that may exist in the atmospheres of these hot giant planets, thereby shedding light on hot Jupiter atmospheric chemistry.

In addition to photometry of exoplanets, the advent of new space-borne instrumentation like the Cosmic Origins Spectrograph[64] on the Hubble Space Telescope opens the possibility of spectroscopic detection of molecular features in the ultraviolet to visible spectral range. Even ground-based spectroscopy of exoplanets may become possible, as astronomers improve the observing and signal-processing techniques needed to isolate the faint exoplanet light[65].

The era of super-Earths begins now

The most exciting of all questions is whether exoplanets host life, so the long-term focus of our interest is on rocky planets in the habitable zones of their stars, where the planet's average temperature permits liquid water. Our own Earth is currently the only site where we know that life exists, and that makes analogues of our Earth the best candidates for hosting life. Unfortunately, an Earth analogue—a twin of our Earth, orbiting in the habitable zone of a solar-type star—will be a difficult world to find and characterize. An Earth analogue lies below the limiting sensitivity of the radial velocity surveys (Fig. 1). The Kepler mission will determine the frequency of occurrence of Earth analogues, and will find many transiting examples (unless they are intrinsically rare). But Kepler surveys stars to relatively faint magnitudes over a limited angular region of the sky, and Earth analogues have a low probability of transiting (about 0.005). Hence the nearest example to be found by Kepler will be distant from us. Eclipse observations (as in Fig. 4) will probably involve too few photons to be useful when applied to an Earth analogue found by Kepler. Eventually, sensitive space-borne astrometric measurements may be able to find Earth analogues orbiting the nearest solar-type stars, and space-borne high-contrast imaging may be able to characterize them. But it will take many years for these missions to be developed and launched, so in the interim we look to other types of habitable planets.

In the search for cosmic life, we naturally turn to super-Earths because their greater mass and larger radii make them easier to find and characterize than Earth analogues. Super-Earths are theorized to vary in composition from solid iron planets, through silicate-dominated planets like our Earth, to water-worlds having bulk compositions that are primarily ice- or water-dominated[66]. The radial velocity surveys have announced about a dozen planets in the mass range of super-Earths, and the K-dwarf star HD 40307 is known to host three of them[67]. One estimate based on the microlensing results[68] is that about one-third of lower main sequence stars host a Neptune-mass or super-Earth-mass planet—this is consistent with a similar estimate based on radial velocity observations[67]. Although the microlensing and radial velocity surveys have given us the tantalizing result that super-Earths are common, especially orbiting lower main sequence stars, these techniques cannot be used to characterize the planets themselves.

Fortunately, transit surveys and transit follow-up observations can both find and characterize nearby super-Earths as they push the domain of their detections down in mass from gas-giants to rocky planets (Fig. 1). The stars nearest to our Sun are predominately low-mass M-dwarf stars. The habitable zone of an M-dwarf is typically ten times closer to the star than is the habitable zone around our Sun, because of the much lower luminosity of M-dwarfs compared to the Sun. Owing to the smaller planet–star separation, planets in the habitable zone of M-dwarfs have a greater probability of transiting than do planets in the habitable zone of solar-type stars. The current frontier in exoplanet science is the push to find and characterize a habitable super-Earth orbiting a nearby M-dwarf star. One reason for this focus is that we can find and characterize such worlds using facilities that are already operable, or in development for flight. Beyond characterizing individual objects, astronomers harbour huge hopes that a large sample of transiting super-Earths with measured masses, radii, semi-major axes, and possibly atmospheric measurements will provide insights on planet formation, migration, interior composition, and evolution.

The next decade

Astronomers have a clear vision of how to discover and characterize super-Earths in the next decade, so as to study the special ones that transit in habitable zones of M-dwarf stars. The ground-based MEarth project[69] is conducting a survey of the nearest 2,000 M-dwarfs to find transiting super-Earths. However, there are more than 10,000 M-dwarfs in the solar neighbourhood (within 35 pc), and a survey of that many stars to the requisite photometric precision will have to be space-based. One such survey that is feasible within the next few years is the proposed Transiting Exoplanet Survey Satellite (TESS) mission. TESS is designed to find a nearby transiting habitable super-Earth, and the best cases found by TESS could be characterized[70] using the James Webb Space Telescope (JWST).

JWST, to be launched in 2014, will be able to characterize the handful of habitable-zone transiting super-Earths we expect to find. It will be able to measure the temperature of such a planet using infrared photometry when the planet is eclipsed[70]. This telescope will also be able to identify the major gases in the planets's atmosphere by observing their absorption during transit, but it will probably not be able to find biosignatures as those are too subtle. The magnitude of the absorption seen during transit (Fig. 2) depends not only on the abundance of the specific molecule being sought, but also on the molecular weight of the atmosphere[71]. Low-molecular-weight atmospheres, for example those that contain significant remnant hydrogen, will have greater pressure scale heights than do high-molecular-weight atmospheres. Greater scale heights increase the area of the absorbing annulus during transit, and provide greater absorption in the transit spectrum[71]. Currently, we do not know the nature of super-Earth atmospheres—that is, whether they are invariably thin (like Earth), and of high molecular weight, or whether they are sometimes much thicker (like Venus) and contain significant hydrogen, either as a remnant from the primordial atmosphere or from outgassing. Hence the fundamental source of uncertainty in this field is the cosmic uncertainty in the nature of the atmospheres on rocky super-Earth worlds.

Super-Earths orbiting in the habitable zones of their host M-stars hold our near-term interest because they are the most accessible planets that have potential to support life. But these planets could be quite different from our own Earth. For example, at small planet–star separations they will be tidally locked, with their sun fixed in their sky at all times, that is, no day–night cycle. Close to their star, the planets may be blasted by ultraviolet radiation from flares that are common on M-dwarf stars, and this would significantly affect[72] the conditions for life on these worlds.

Decades beyond 2020

The astronomical community is developing technology for new space missions to find and characterize true Earth analogues in future decades. One concept[73] calls for a flagship successor to the Hubble Space Telescope: this telescope would operate at ultraviolet–visible wavelengths, and be equipped with an external occulter. It would act part-time as a planet finder, and (while the occulter was moving to another target) part-time as a general observatory. An exciting alternative is to launch an external occulter[4] to be used in conjunction with the JWST, enabling a Terrestrial Planet Finder to happen within one decade. Whatever concept becomes reality, light from a truly Earth-like world is within the grasp of our instruments in the foreseeable future.

1. Mayor, M. & Queloz, D. A. Jupiter-mass companion to a solar-type star. *Nature* **368,** 355–359 (1995).
   The first detection of a planet orbiting a solar-type star.
2. Marcy, G. *et al.* Exoplanet properties from Lick, Keck and AAT. *Phys. Scripta T* **130,** 014001 (2008).
3. Trauger, J. T. & Traub, W. A. A laboratory demonstration of the capability to image an Earth-like extrasolar planet. *Nature* **446,** 771–773 (2007).
4. Cash, W. Detection of Earth-like planets around nearby stars using a petal-shaped occulter. *Nature* **442,** 51–53 (2006).
   Proposes a promising and practical method for imaging Earth-analogue exoplanets.
5. Kalas, P. *et al.* Optical images of an extrasolar planet 25 light-years from Earth. *Science* **322,** 1345–1347 (2008).
6. Marois, C. *et al.* Direct imaging of multiple planets orbiting the star HR 8799. *Science* **322,** 1348–1352 (2008).
7. Marley, M. S., Fortney, J. J., Hubickyj, O., Bodenheimer, P. & Lissauer, J. J. On the luminosity of young Jupiters. *Astrophys. J.* **655,** 541–549 (2007).
8. Chiang, E., Kite, E., Kalas, P., Graham, J. R. & Clampin, M. Fomalhaut's debris disk and planet: constraining the mass of Fomalhaut b from disk morphology. *Astrophys. J.* **693,** 734–749 (2009).
9. Charbonneau, D., Brown, T. M., Burrows, A. & Laughlin, G. in *Protostars and Planets V* (eds Reipurth, B., Jewitt, D. & Keil, K.) 701–716 (Univ. Arizona Press, 2007).
10. Charbonneau, D., Brown, T. M., Latham, D. & Mayor, M. Detection of planetary transits across a Sun-like star. *Astrophys. J.* **529,** L45–L48 (2000).
    Reports the first full transit of a planet across a Sun-like star.
11. Henry, G. W., Marcy, G. W., Butler, R. P. & Vogt, S. S. A transiting "51-Peg-like" planet. *Astrophys. J.* **529,** L41–L44 (2000).
12. Guillot, T. & Showman, A. P. Evolution of "51-Pegasus-b-like" planets. *Astron. Astrophys.* **385,** 156–165 (2002).
13. Miller, N., Fortney, J. J. & Jackson, B. Inflating and deflating hot Jupiters: coupled tidal and thermal evolution of known transiting planets. *Astrophys. J.* (submitted); preprint at ⟨http://arxiv.org/abs/0907.1268⟩ (2009).
14. Deming, D., Seager, S., Richardson, L. J. & Harrington, J. Infrared radiation from an extrasolar planet. *Nature* **434,** 740–743 (2005).
    This paper and ref. 24 simultaneously report the first detection of light from an exoplanet.
15. Knutson, H. A., Charbonneau, D., Burrows, A., O'Donovan, F. T. & Mandushev, G. Detection of a temperature inversion in the broadband infrared emission spectrum of TrES-4, 2009. *Astrophys. J.* **691,** 866–874 (2009).
16. Chabrier, G. & Baraffe, I. Heat transport in giant (exo)planets: a new perspective. *Astrophys. J.* **661,** L81–L84 (2007).
17. Burke, C. J. *et al.* XO-5b: a transiting Jupiter-sized planet with a 4-day period. *Astrophys. J.* **686,** 1331–1340 (2008).
18. Bakos, G. A. *et al.* HAT-P-10b: a light and moderately hot Jupiter transiting a K-dwarf. *Astrophys. J.* **696,** 1950–1955 (2009).
19. West, R. G. *et al.* The low density transiting exoplanet WASP-15b. *Astron. J.* **137,** 4834–4836 (2009).
20. Gillon, M. *et al.* Detection of transits of the nearby hot Neptune GJ436b. *Astron. Astrophys.* **472,** L13–L16 (2007).
    GJ436b is the first transiting planet significantly smaller than Jupiter.
21. Bakos, G. A. *et al.* HAT-P-11b: a super-Neptune planet transiting a bright K star in the Kepler field. *Astrophys. J.* (in the press); preprint at ⟨http://arxiv.org/abs/0901.0282⟩ (2009).
22. Queloz, D. *et al.* The CoRoT-7 planetary system: two orbiting super-Earths. *Astron. Astrophys.* (in the press).
    CoRoT-7b is the first transiting super-Earth exoplanet.
23. Borucki, W. *et al.* in *Transiting Planets* (eds Pont, F., Sasselov, D. & Holman, M. J.) 289–299 (IAU Symp. 253, Cambridge Univ. Press, 2009).
24. Charbonneau, D. *et al.* Detection of thermal emission from an extrasolar planet. *Astrophys. J.* **626,** 523–529 (2005).
    This paper and ref. 14 simultaneously report the first detection of light from an exoplanet.
25. Deming, D., Harrington, J., Seager, S. & Richardson, L. J. Strong infrared emission from the extrasolar planet HD 189733b. *Astrophys. J.* **644,** 560–564 (2006).
26. Charbonneau, D. *et al.* The broadband infrared emission spectrum of the exoplanet HD 189733b. *Astrophys. J.* **686,** 1341–1348 (2008).
27. Knutson, H. A. *et al.* A map of the day-night contrast of the extrasolar planet HD189733b. *Nature* **447,** 183–186 (2007).
    A long series of Spitzer observations enabled a longitudinal temperature map of an exoplanet for the first time.

28. Harrington, J., Luszcz, S., Seager, S., Deming, D. & Richardson, L. J. The hottest planet. *Nature* **447,** 691–693 (2007).
29. Knutson, H. A., Charbonneau, D., Allen, L. E., Burrows, A. & Megeath, S. T. The 3.6–8.0 micron broadband emission spectrum of HD 209458b: evidence for an atmospheric temperature inversion. *Astrophys. J.* **673,** 526–531 (2008).
Many giant exoplanets have temperature inversions in their atmospheres, and HD209458b is the archetype for this important phenomenon.
30. Machalek, P. *et al.* Thermal emission of exoplanet XO-1b. *Astrophys. J.* **684,** 1427–1432 (2008).
31. Machalek, P., McCullough, P. R., Burrows, A., Burke, C. J. & Hora, J. L. Detection of thermal emission of XO-2b: evidence for a weak temperature inversion. *Astrophys. J.* **701,** 514–520 (2009).
32. Seager, S. & Sasselov, D. D. Theoretical transmission spectra during extrasolar giant planet transits. *Astrophys. J.* **537,** 916–921 (2000).
33. Sudarsky, D., Burrows, A. & Hubeny, I. Theoretical spectra and atmospheres of extrasolar giant planets. *Astrophys. J.* **588,** 1121–1148 (2003).
34. Seager, S. *et al.* On the dayside thermal emission of hot Jupiters. *Astrophys. J.* **632,** 1122–1131 (2005).
35. Rowe, J. F. *et al.* The very low albedo of an extrasolar planet: MOST space-based photometry of HD 209458. *Astrophys. J.* **689,** 1345–1353 (2008).
36. Alonso, R. *et al.* The secondary eclipse of the transiting exoplanet CoRoT-2b. *Astron. Astrophys.* **501,** L23–L26 (2009).
37. Marley, M. S., Gelino, C., Stephens, D., Lunine, J. I. & Freedman, R. Reflected spectra and albedos of extrasolar giant planets. I. Clear and cloudy atmospheres. *Astrophys. J.* **513,** 879–893 (1999).
38. Seager, S., Whitney, B. A. & Sasselov, D. D. Photometric light curves and polarization of close-in extrasolar giant planets. *Astrophys. J.* **540,** 504–520 (2000).
39. Richardson, L. J., Deming, D., Horning, K., Seager, S. & Harrington, J. A spectrum of an extrasolar planet. *Nature* **445,** 892–895 (2007).
40. Grillmair, C. J. *et al.* A Spitzer spectrum of the exoplanet HD 189733b. *Astrophys. J.* **658,** L115–L118 (2007).
41. Grillmair, C. J. *et al.* Strong water absorption in the dayside emission spectrum of the planet HD189733b. *Nature* **456,** 767–769 (2008).
Reports the unequivocal detection of water vapour in the spectrum of an exoplanet.
42. Swain, M. R., Vasisht, G. & Tinetti, G. The presence of methane in the atmosphere of an extrasolar planet. *Nature* **452,** 329–331 (2008).
43. Swain, M. R. *et al.* Molecular signatures in the near-infrared dayside spectrum of HD 189733b. *Astrophys. J.* **690,** L114–L117 (2009).
44. Vidal-Madjar, A. *et al.* An extended upper atmosphere around the extrasolar planet HD209458b. *Nature* **422,** 143–146 (2003).
45. Charbonneau, D., Brown, T. M., Noyes, R. W. & Gilliland, R. L. Detection of an extrasolar planet atmosphere. *Astrophys. J.* **568,** 377–384 (2002).
46. Harrington, J. *et al.* The phase-dependent infrared brightness of the extrasolar planet upsilon Andromedae b. *Science* **314,** 623–626 (2006).
47. Cowan, N. B., Agol, E. & Charbonneau, D. Hot nights on extrasolar planets: mid-infrared phase variations of hot Jupiters. *Mon. Not. R. Astron. Soc.* **379,** 641–646 (2007).
48. Knutson, H. A. *et al.* The 8-micron phase variation of the hot Saturn HD 149026b. *Astrophys. J.* **703,** 769–784 (2009).
49. Showman, A. P., Menou, K. & Cho, J. Y.-K. in *Extreme Solar Systems* (eds Fischer, D., Rasio, F. A., Thorsett, S. E. & Wolszczan, A.) 419 (ASP Conf. Ser. Vol. 398, Astronomical Society of the Pacific, 2008).
50. Menou, K. & Rauscher, E. Atmospheric circulation of hot Jupiters: a shallow three-dimensional model. *Astrophys. J.* **700,** 887–897 (2009).
51. Hubeny, I., Burrows, A. & Sudarsky, D. A possible bifurcation in atmospheres of strongly irradiated stars and planets. *Astrophys. J.* **594,** 1011–1018 (2003).
52. Fortney, J. J., Lodders, K., Marley, M. S. & Freedman, R. S. A unified theory for the atmospheres of the hot and very hot Jupiters: two classes of irradiated atmospheres. *Astrophys. J.* **678,** 1419–1435 (2008).
53. Zahnle, K., Marley, M. S., Freedman, R. S., Lodders, K. & Fortney, J. J. Atmospheric sulphur photochemistry on hot Jupiters. *Astrophys. J.* **701,** L20–L24 (2009).
54. Laughlin, G. *et al.* Rapid heating of the atmosphere of an extrasolar planet. *Nature* **457,** 562–564 (2009).
55. Winn, J. *et al.* HAT-P-7: a retrograde or polar orbit, and a third body. *Astrophys. J.* **703,** L99–L103 (2009).
56. Fabrycky, D. & Tremaine, S. Shrinking binary and planetary orbits by Kozai cycles with tidal friction. *Astrophys. J.* **669,** 1298–1315 (2007).
57. Yu, Q. & Tremaine, S. Resonant capture by inward-migrating planets. *Astron. J.* **121,** 1736–1740 (2001).
58. Lopez-Morales, M. & Seager, S. Thermal emission from transiting very hot Jupiters: prospects for ground-based detection at optical wavelengths. *Astrophys. J.* **667,** L191–L194 (2007).
59. Snellen, I. A. G., de Mooij, E. J. W. & Albrecht, S. The changing phases of extrasolar planet CoRoT-1b. *Nature* **459,** 543–545 (2009).
60. Borucki, W. J. *et al.* Kepler's optical phase curve of the exoplanet HAT-P-7b. *Science* **325,** 709 (2009).
Kepler's spectacular photometric precision means that transiting Earth-analogue exoplanets are now within reach.
61. Ballard, S. *et al.* A search for additional planets in the NASA EPOXI observations of GJ 436. *Astrophys. J.* (submitted); preprint at ⟨http://arxiv.org/abs/0909.2875⟩ (2009).
62. de Mooij, E. J. W. & Snellen, I. A. G. Ground-based K-band detection of thermal emission from the exoplanet TrES-3b. *Astron. Astrophys.* **493,** L35–L38 (2009).
63. Gillon, M. *et al.* VLT transit and occultation photometry for the bloated planet CoRoT-1b. *Astron. Astrophys.* (in the press).
64. Froning, C. & Greene, J. C. The cosmic origins spectrograph: capabilities and pre-launch performance. *Astrophys. Space Sci.* **320,** 181–185 (2009).
65. Barnes, J. R. A search for molecules in the atmosphere of HD189733b. *Mon. Not. R. Astron. Soc.* (in the press); preprint at ⟨http://arxiv.org/abs/0909.2510⟩ (2009).
66. Seager, S., Kuchner, M., Hier-Majumder, C. A. & Militzer, B. Mass-radius relationship for solid exoplanets. *Astrophys. J.* **669,** 1279–1297 (2007).
Predicts the global properties of super-Earth exoplanets.
67. Mayor, M. *et al.* The HARPS search for southern extrasolar planets XIII. A planetary system with three super-Earths. *Astron. Astrophys.* **493,** 639–644 (2009).
Concludes that super-Earth exoplanets are common around stars similar to our Sun.
68. Gould, A. *et al.* Microlens OGLE-2005-BLG-169 implies that cool Neptune-like planets are common. *Astrophys. J.* **644,** L37–L40 (2006).
69. Nutzman, P. & Charbonneau, D. Design considerations for a ground-based transit search for habitable planets orbiting M-dwarfs. *Publ. Astron. Soc. Pacif.* **120,** 317–327 (2007).
70. Deming, D. *et al.* Discovery and characterization of transiting superEarths using an all-sky transit survey, and follow-up by the James Webb Space Telescope. *Publ. Astron. Soc. Pacif.* **121,** 952–967 (2009).
71. Miller-Ricci, E., Seager, S. & Sasselov, D. The atmospheric signatures of super-Earths: how to distinguish between hydrogen-rich and hydrogen-poor atmospheres. *Astrophys. J.* **690,** 1056–1067 (2009).
72. Tarter, J. C. *et al.* A reappraisal of the habitability of planets around M-dwarf stars. *Astrobiology* **7,** 30–65 (2007).
73. Spergel, D. N. *et al.* THEIA: Telescope for Habitable Exoplanets and Interstellar/Intergalactic Astronomy. *Bull. Am. Astron. Soc.* 213, abstr. 458.04 (2009).
74. Fortney, J. J., Marley, M. S. & Barnes, J. W. Planetary radii across five orders of magnitude in mass and stellar insolation: application to transits. *Astrophys. J.* **659,** 1661–1672 (2007).
75. Burrows, A., Budaj, J. & Hubeny, I. Theoretical spectra and light curves of close-in extrasolar planets and comparison with data. *Astrophys. J.* **678,** 1436–1457 (2008).

Acknowledgements This work is based in part on observations made with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. Support for this work was provided directly by NASA, and by NASA through an award issued by JPL/Caltech.

Author Contributions Both authors contributed equally to this work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence should be addressed to D.D. (Leo.D.Deming@nasa.gov).

REVIEWS

Human genetics illuminates the paths to metabolic disease

Stephen O'Rahilly[1]

Metabolic diseases represent a growing threat to world-wide public health. In general, these disorders result from the interaction of heritable factors with environmental influences. Here, I will focus on two important metabolic disorders, namely type 2 diabetes and obesity, and explore the extent to which human molecular genetic research has illuminated our understanding of their underlying pathophysiological mechanisms.

Definitive measures of the heritability of a human trait or disease are unattainable as the precise extent to which inherited factors determine inter-individual differences in risk varies between populations and over time. When considered as quantitative traits, measures of 'fatness' are highly heritable with about 70% of inter-individual difference in indices of adiposity being attributable to genetic factors[1]. Twin studies suggest that although glucose tolerance is highly heritable, non-genetic factors also have a role, particularly in determining the timing of transition to frank diabetes[2]. It is clear that there has been, in many countries, a rapid recent increase in both obesity and the type 2 diabetes, especially in children, and changes over this time frame are unlikely to be driven by genetic alterations[3]. Nevertheless, even in the midst of the 'obesity epidemic', genetic factors still seem to have a major role in determining who becomes obese[4] and/or develop diabetes.

Understanding the heritable component of metabolic disease

Although simple curiosity remains a noble impetus to scientific effort, the enormous investment in human genomics has been driven by the promise that increasing genetic knowledge would translate into improved tools for the treatment and prevention of disease. This promise currently sits uncomfortably with the paucity of novel, safe and effective treatments emerging from the pharmaceutical sector. In engaging with the public and our funders we need to emphasize the long-term view of the war against complex chronic metabolic disease, while celebrating any successes we achieve along the way. In the case of obesity and diabetes we already have compelling examples where genetic knowledge has improved the health and well being of some people[5,6]. In the future, increasing knowledge of the genetic architecture of metabolic disease is likely to deliver tangible benefits to human health. First, by discovering and validating key nodal points in the control of key elements of metabolic homeostasis, human genetics is likely to inform decisions regarding the selection of molecular targets for novel therapeutics. Second, we can predict that the reliable dissection of genetic and pathophysiological heterogeneity within metabolic diseases that are currently artificially grouped as single entities should lead to improvements in personalized diagnosis, prognostication, therapy and prevention. However, it is likely that such impacts will occur in a gradual and stepwise manner.

Type 2 diabetes

Diabetes is a condition defined by a state of chronic elevation of plasma glucose levels, the adverse impact of which occurs predominantly, but not exclusively, through its effects on the health of small and large blood vessels. Much effort has gone into the classification of diabetes. In simple terms, type 1 diabetes results from an autoimmune destruction of the insulin-producing pancreatic beta cells, and forms of diabetes that result from defined causes such as monogenic disorders of insulin secretion or action, or as a secondary consequence of acquired pancreatic, endocrine or other disorders are classified individually according to their primary cause[7]. However, with our current knowledge, these subtypes together only account for $<$10% of all cases of diabetes[7]. The term 'type 2 diabetes' is currently used for the remainder, which is very unlikely to be a homogenous entity.

The view of type 2 diabetes in the 'pre-molecular genetic' era. It has long been clear that (1) people with type 2 diabetes are very often obese or overweight; (2) the majority of patients with type 2 diabetes have an impaired metabolic response to administered or endogenous insulin; and (3) obesity, even in non-diabetic individuals, tends to lead to a state of insulin resistance[8]. Beginning in the 1960s, when it first became possible to measure plasma levels of insulin, a vast body of pathophysiological investigation of people with diabetes and 'prediabetes' led to a general (if not universal) consensus that the inherited defect predisposing to type 2 diabetes was likely to involve a primary defect in insulin action[9]. As the major tissue responsible for glucose disposal after a meal, skeletal muscle was widely considered to be the location at which such a defect would primarily express itself[10]. Thus, insulin resistance, exacerbated by obesity, would result in chronic overwork of pancreatic beta cells, ultimately leading to their 'exhaustion' and decompensation to a hyperglycaemic state. However, throughout this period, some investigators continued to draw attention to the fact that normoglycaemic 'at risk' individuals seemed to have both quantitative and qualitative defects in pancreatic beta-cell function that indicated the existence of an intrinsic problem with islet function[11].

Lessons from monogenic forms of diabetes. Over the past 20 years or so, arguably the major contribution of molecular genetics to 'non-type-1 diabetes' has been the identification of a range of single gene disorders that result in chronic hyperglycaemia (Table 1). The elucidation of the precise molecular basis for a number of these conditions has been illuminating in terms of the basic understanding of critical components in the control of human glucose homeostasis and in the lessons they have taught us that are relevant to our concepts of the pathophysiology of common forms of diabetes. Importantly, these discoveries have also brought some practical improvements in diagnosis and management.

[1]University of Cambridge Metabolic Research Laboratories, Institute of Metabolic Science, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK.

Table 1 | Monogenic disorders of glucose tolerance

Gene	Functional defect or consequence	Reference
Inherited disorders largely impacting on insulin secretion		
HNF4a *HNF1A* *IPF1* (also called *PDX1*) *HNF1B* *NEUROD1*	Defects in transcription factors responsible for beta-cell development, maintenance, function or survival	14 15 16 17 18
GCK (Glucokinase) *ABCC8* (sulphonylurea receptor) *KCNJ11* (Kir6.2 potassium channel) Mitochondrial 3243 variant	Defects in the glucose sensing/ metabolism/secretory function of pancreatic beta cells	13 19 20 49
Chromosome 6q22-6q23 *ZFP57*	Imprinting disorders causing transient neonatal diabetes	79 80
EIF2AK3 *WFS1* *INS* *CEL*	Defects leading to increased beta-cell death; increased ER stress (*WFS1* and *INS*)	53 50 21 81
Inherited disorders largely impacting on insulin action		
*AGPAT2**	Defects in adipocyte triglyceride synthesis	31
*PPARG** *BSCL2* (Seipin)*	Defects in adipocyte development	28 30
LMNA (lamin A/C)* *CAV1* (Caveolin 1)*	Mechanism unclear	20 32
*CIDEC**	Dysregulated lipolysis	33
INSR† *AKT2*† *TBC1D4*† (also known as *AS160*)		23 25 26
ALMS1‡	Alstrom's syndrome; mechanism unclear	82,83

Examples of mechanisms whereby mutations in single human genes lead to disorders characterized by impaired glucose tolerance. This list is not comprehensive. For a fuller review see ref. 84.
* Primary lipodystrophies.
† Disorders of insulin signalling.
‡ Pleiotropic syndromes associated with severe insulin resistance.

Inherited disorders of insulin secretion. In the early 1970s clinicians reported a form of 'non-type-1' diabetes that presented in young adult life, appeared to be highly familial, and did not require insulin therapy for its control. They termed this 'maturity onset diabetes of the young' (MODY) and it subsequently became clear that, at least in some families, inheritance of this disorder followed an autosomal-dominant pattern[12]. Physiological studies suggested that, in most cases, MODY resulted from a defect in insulin secretion, rather than action[12]. The first MODY gene to be identified encodes glucokinase, an enzyme involved in the sensing of glucose by the pancreatic beta cell[13]. Subsequently, the application of linkage analysis identified mutations in a number of transcription factors, several in the HNF family of basic-loop-helix transcription factors, to be the cause of other subtypes of MODY[14–18]. More recently it has been realized that some children who presented with diabetes very early in life did not have 'typical' type 1 diabetes but rather had monogenic forms of neonatal diabetes arising from mutations that result in impaired beta-cell development or function or increased beta-cell destruction[19–21]. In the case of MODY, the best estimates are that the aetiology of >80% of cases of early-onset, autosomal-dominant, familial hyperglycaemia is now identifiable[22]. The vast majority of genetic causes identified so far involve a primary deficiency in insulin secretion[22].

Inherited disorders of insulin action. The normal control of glucose metabolism requires not only a normal secretion of insulin but also its normal action in its key target tissues including muscle, liver and fat. It might, therefore, be considered surprising that when monogenic forms of diabetes are parsed, insulin secretory defects seem to be grossly overrepresented. Why are genetic disorders of insulin action not more prominent? Studies of familial forms of inherited insulin resistance provide at least some of the answer to this. Mutations in the human insulin receptor gene were the first form of primary genetic disorder of human insulin action to be described[23]. Intriguingly, not all patients carrying such mutations present with diabetes, as the response of the healthy pancreas to such intrinsic insulin resistance is to expand islet beta-cell mass and hypersecrete enormous quantities of insulin[24]. In some individuals at least, this could maintain normoglycaemia for many decades, although most subjects do finally decompensate to develop diabetes, which is then extremely difficult to treat[24]. The capacity of compensatory hyperinsulinaemia to ward off diabetes for many years explains why, in contrast to single gene disorders affecting insulin secretion, single gene disorders impacting on insulin action frequently do not manifest themselves as diabetes until rather late in life. So far, other than mutations in the insulin receptor, only a single human family (with a mutation in *AKT2*) has been reported in which a primary defect in the insulin signal transduction pathway leads to a clearly monogenic form of diabetes[25]. We have recently described a human family where a nonsense mutation in *TBC1D4*, which encodes a Rab-GAP protein involved in glucose transporter translocation in skeletal muscle and fat, causes severe selective post-prandial insulin resistance[26]. However, the compensatory capacity of the beta cell is such that diabetes does not necessarily ensue.

Although monogenic disorders directly impairing the action of insulin have been hard to find, there is another group of single gene disorders leading to insulin-resistant forms of diabetes where considerable progress has been made. The failure to develop, or adequately store triglyceride in, adipose tissue leads to ectopic deposition of stored lipid in muscle, liver, pancreas and elsewhere, and this phenomenon, through mechanisms that are being aggressively explored, severely impairs insulin action in those tissues[27]. Although compensatory hyperinsulinaemia prevents diabetes for a time, diabetes frequently supervenes. The genetic defects underlying several forms of recessive and dominantly inherited lipodystrophic disorders have been identified[28–33]. Several of these causative genes are uniquely expressed in adipocytes and thus demonstrate that adipose tissue 'failure' alone can initiate a pathophysiological cascade that can lead to an insulin-resistant form of human diabetes.

Lessons from genetic studies in 'common' type 2 diabetes. Much effort has been expended in the attempt to identify common genetic variants which underpin 'common' type 2 diabetes (Table 2). However, it is only recently, when the key challenges of adequate

Table 2 | Common genetic variants associated with type 2 diabetes

Chromosome	Nearby genes	How identified	Reference
1	*NOTCH2*	GWAS meta-analysis	85
2	*THADA*	GWAS meta-analysis	85
2	*IRS1*	GWAS	77
3	*ADAMTS9*	GWAS meta-analysis	85
3	*PPARG*	Candidate gene	86
3	*IGF2BP2**	GWAS	87,88,89
4	*WFS1*	Candidate gene	51
6	*CDKAL1**	GWAS	87,88,89,90
7	*JAZF1*	GWAS meta-analysis	85
8	*SLC30A8**	GWAS	91
9	*CDKN2A/B**	GWAS	87,88,89
10	*CDC123, CAMK1D*	GWAS meta-analysis	85
10	*HHEX**, *KIF11, IDE*	GWAS	91
10	*TCF7L2**	Linkage analysis/region-wide genotyping	92
11	*KCNJ11**	Candidate gene	93
11	*MTNR1B**	GWAS	41
12	*TSPAN8, LGR5*	GWAS meta-analysis	85
16	*FTO*	GWAS	69
17	*HNF1B*	Candidate gene	94

Susceptibility loci associated with type 2 diabetes have been described in detail in several recent reviews (for example, ref. 34). Increasing numbers of SNPs replicably associated with type 2 diabetes are rapidly emerging and it is unlikely that this table will be up to date at the time of publication. GWAS, genome-wide association study.
* SNPs associated with decreased insulin secretion in non-diabetic humans[35–39].

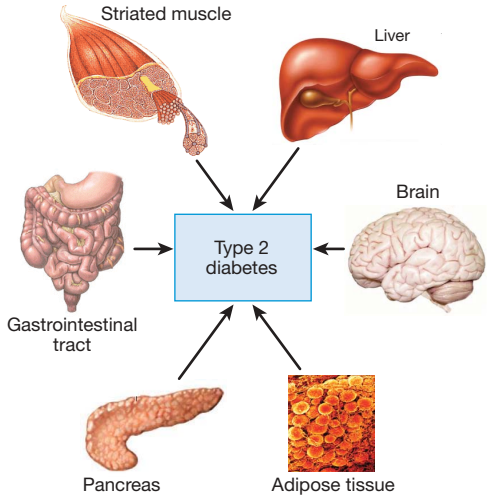
sample size and genome-wide coverage of genetic variation have been met, that reliable and reproducible information has finally emerged. There are now at least 19 common alleles (present in at least 1% of the populations that have been studied) that are generally accepted as being truly associated with type 2 diabetes[34]. More are very likely to emerge. A major scientific challenge over the forthcoming years will be to put mechanistic flesh on the bare bones of these associations. Although a minority of these diabetes single nucleotide polymorphisms (SNPs) are actually known to affect the structure or expression of a gene product with a credible link to glucose metabolism, the majority are not. It is likely that the physical location of these SNPs in the genome will give some clue to their ultimate biological effect, with the most closely co-located functional gene most likely to be the culprit; however, 'guilt by proximity' cannot be universally assumed. Nevertheless, it is striking that a large number of diabetes SNPs are close to genes expressed highly in the adult or developing pancreas and many have been shown to be associated with reduced beta-cell dysfunction in non-diabetic subjects[35–39]. Investigators have also used genome-wide approaches to examine genetic determinants of fasting glucose levels as a quantitative trait and it is intriguing that there is only limited overlap with 'diabetes' genes[40–42].

Human genetics informs pathophysiology of type 2 diabetes. The results from studies of monogenic diabetes remind us of the salient facts that the normal pancreatic beta cell can upregulate its function and maintain normoglycaemia for some considerable time in the face of severe insulin resistance, but that even a modest inherent defect in beta-cell function leads to hyperglycaemia[22]. That said, it is still formally possible that the majority of genes predisposing to common type 2 diabetes would exert their effect on insulin resistance with 'normal' beta cells simply failing in response to decades of compensatory overwork. The recent genome-wide association studies suggest that this simple scenario is likely to be incorrect. It seems more probable that a substantial part of the inherent susceptibility of an individual to develop type 2 diabetes in mid-life relates to the extent to which pancreatic beta-cell function can be maintained (Box 1). Looking back at the 'pre-molecular genetic' literature, could we have predicted that? There actually was a 'contrarian' tendency, much of it emanating from Europe, which in the 1970s and 1980s tried to draw attention to the presence, in non-diabetic people at high risk of later type 2 diabetes, of subtle quantitative and qualitative abnormalities of insulin secretion[11,43].

This reformulation of the general view of type 2 diabetes does not mean that insulin resistance is not important in its aetiology. Indeed most behavioural or pharmacological interventions that are proven to delay the onset of type 2 diabetes have their major impact on insulin sensitivity[44]. However, rather than viewing such interventions as targeting the sole and fundamental mechanism of type 2 diabetes, we should now perhaps better see them as protecting the inherently vulnerable pancreatic beta cell from overwork. That said, even in the absence of hyperglycaemia, insulin resistance is strongly associated with, and may even have a causative role in, important and highly prevalent conditions such as atherosclerosis, dyslipidaemia, hypertension, non-alcoholic steatohepatitis and polycystic ovarian syndrome, and thus improving our understanding of its aetiology is a biomedical priority.

For several decades diabetes researchers tended to focus their research on the pancreatic beta cell or on the two principal targets controlling glucose production and utilization in the whole body, namely liver and skeletal muscle. As only a small fraction of ingested glucose ended up in adipose tissue, the latter was, for some time, relatively ignored. The dawning realization that the healthy handling of carbohydrate metabolism was crucially interconnected with and dependent on lipid metabolism led to the emergence of the adipocyte as the 'fourth musketeer' of diabetes[27,45]. The molecular genetic era has clearly shown us that genetic defects solely impairing the development and/or function of adipose tissue (as in the hereditary lipodystrophies) are sufficient to result in the early and inevitable appearance of diabetes, providing compelling proof of the importance of that previously neglected organ.

Box 1 | The complex landscape of type 2 diabetes



Monogenic causes of diabetes impact mostly on pancreatic islet function, and many 'common' type 2 diabetes SNPs map close to genes expressed in islets and some are associated with beta-cell dysfunction in non-diabetic subjects. These findings have shifted the 'model' of type 2 diabetes from one where intrinsically normal pancreatic islets become 'exhausted' as a result of prolonged attempts to compensate for a primary defect in insulin action to one where intrinsic genetic variability in islet function is a major determinant of the susceptibility to develop diabetes.

Unlike muscle and liver, adipose tissue is not a major site of either glucose disposal or production. However, monogenic disorders of adipose tissue development/function lead to severe insulin resistance/diabetes as a result of 'fat failure' and diversion of nutrient delivery to muscle and liver, which impairs insulin action in these tissues through mechanisms that remain to be determined conclusively.

The brain clearly has a leading role in the control of energy balance, and obesity is a major risk factor for type 2 diabetes. In animal models, physiological/pharmacological manipulations in the brain influence glucose metabolism independent of energy homeostasis. So far, human genetic variants associated with both obesity and type 2 diabetes seem to exert most of their effects on glucose tolerance through an effect on adiposity (which itself increases insulin resistance in muscle and liver).

Muscle and liver are the major sites of insulin-regulated glucose disposal/production. Although insulin resistance, especially in muscle, is a key, early feature of type 2 diabetes and is heritable, surprisingly few 'diabetes genetic hits' primarily affect these tissues. Why is this so? First, environmental factors (diet, exercise, and so on) have major, highly labile impacts on these tissues. Insulin resistance is therefore an unstable phenotype and genetic influences are hard to 'pin down'. Second, effects of inherited variation in insulin action on glucose metabolism are masked by the capacity of normal pancreas to compensate.

Physiological and pharmacological studies indicate a powerful influence of hormonal products of entero-endocrine cells on glucose homeostasis. Human genetics has not, as yet, convincingly contributed to the notion that variation in entero-endocrine cell function has a role in determining susceptibility to diabetes. However, *TCF7L2*, which is at the site of the first discovered and most significant SNP associated with 'common' type 2 diabetes, is expressed in entero-endocrine cells and might influence their function.

Research efforts into possible pathogenic mechanisms in type 2 diabetes continue to be intense, with novel and potentially unifying hypotheses continually emerging. Of these, I will briefly consider three—namely mitochondrial dysfunction[46], endoplasmic reticulum (ER) stress[47] and inflammation[48]—and explore the extent to which human molecular genetics has supported the potential aetiological importance of these processes in human diabetes. Certain mutations in the mitochondrial genome are consistently associated with maternally inherited forms of human diabetes, and those mutations seem to cause

diabetes largely by impairing insulin secretion rather than insulin action[49]; however, there is currently little evidence to implicate common variation in the mitochondrial genome, or indeed in nuclear encoded/mitochondrially expressed genes in type 2 diabetes. *WFS1* was found to be mutated in Wolfram syndrome[50], a pleiotropic recessive disorder associated with diabetes. Subsequently, mutations in *WFS1* were found in children with non-syndromic forms of insulinopaenic diabetes and common variants in *WFS1* influence pancreatic beta-cell function and type 2 diabetes risk[51]. WFS1 has been implicated in the control of ER stress responses[52], and some other monogenic forms of diabetes (*EIF2AK3*[53] and *INS*[46]) have also been found to impair pancreatic beta-cell function by interfering with, or overwhelming, normal ER stress responses. As with mitochondrial dysfunction the principal impact of these disorders is on insulin secretion with little evidence yet for an effect on insulin resistance. There is good evidence that patients with pre-diabetes and type 2 diabetes have low levels of activation of inflammatory pathways[48], but the evidence that such activation is aetiologically involved in the disease is currently lacking. In contrast with, for example, type 1 diabetes or inflammatory bowel disease[54], the genome-wide studies of type 2 diabetes do not suggest an inflammatory 'genetic signature'. Results in genome-wide association studies of insulin-resistant phenotypes are awaited with interest.

Clinical implications of genetic discoveries in diabetes. At present, the information that has emerged regarding the genetic basis of common type 2 diabetes has not penetrated into clinical practice. Even taking all the known diabetes risk SNPs together adds only marginally more predictive power to that provided by conventional risk factors[55,56]. In contrast, the work in MODY has led to real clinical utility and patient benefit[22]. It is now clear that patients with glucokinase mutations have life-long, stable modest hyperglycaemia and do not need aggressive glucose-lowering therapy, whereas patients with other forms of MODY tend to progress and eventually require insulin therapy[22]. Notably, patients with HNF1-α mutations seem to be supersensitive to sulphonylurea drugs[57], a fact that can be used to harness these drugs to patient benefit while avoiding dangerous hypoglycaemia. Finally, and most strikingly, a rare group of patients with mutations that result in the voltage-gated potassium channel of the pancreatic beta cell being constitutively open[19,20] and who consequently develop insulin-dependent diabetes in the neonatal period have been shown to respond to sulphonylurea drugs[5]. Such patients can stop insulin injections and achieve improved control with an oral medication, resulting in a dramatic improvement in their quality of life and likely long-term health.

Obesity

Obesity is most simply defined as a state in which the total amount of triglyceride stored in adipose tissue is abnormally increased. It is strongly associated with a wide variety of adverse health outcomes, including diabetes, vascular disease and certain cancers[58]. Obesity results from a chronic, positive imbalance between energy intake and energy expenditure (Box 2). Intuitively, it seems that simple measurements of the components of these two sides of the energy equation should be readily able to identify the major contributor to obesity in any single affected individual. In practice, however, the accrual of excess fat mass usually occurs gradually and the daily imbalance of energy required to result in obesity is so small as to seriously challenge the resolution of existing measures of energy intake or expenditure. Additionally, obesity is highly susceptible to the 'Hawthorne' effect, with individuals who know they are being studied consciously or unconsciously altering their behaviour. The fact that adiposity is a highly heritable trait does, however, provide the opportunity to use modern molecular genetics to obtain mechanistic insights that were previously unobtainable. If we could find variants in genes with known or at least tractable functions that are unequivocally associated with obesity we might start to be able to build up a picture of what sorts of biological factors determine why, in the face of a highly 'obesogenic' environment, some people are susceptible to obesity whereas others remain lean.

The view of obesity in the 'pre-molecular genetic' era. Much effort was expended in exploring the hypothesis that obese people had a reduced basal metabolic rate (BMR). This was convincingly demonstrated not to be the case[59]. Indeed obesity was generally found to be associated with increased BMR because the parallel increase in lean mass is the main determinant of BMR. Obese people generally reported a low daily caloric intake, certainly insufficient to explain their obesity[60]. Several studies showed that obese people tended to underestimate consistently their daily caloric intake[60]. This challenging set of observations did not help to make human obesity a particularly attractive area in which to pursue a scientific or medical career.

Lessons from monogenic disorders. It had long been known that human obesity could result from a disorder in a single gene. Clinically defined genetic syndromes such as Prader–Willi and Bardet–Biedl were strongly associated with obesity and a voracious appetite but the coexistence of developmental delay and more generalized brain dysfunction complicated interpretation. However, in the mid-1990s the first human single genetic defects were found that led to severe obesity in the absence of developmental delay[61,62]. These discoveries

Table 3 | Monogenic disorders leading to human obesity

Disorders without generalized disturbance of higher nervous system functions

Gene	Encoded product	Comment	Reference
LEP	Leptin	Adipocyte-derived hormone	61
LEPR	Leptin receptor	Receptor for adipocyte-derived hormone	95
POMC	Pro-opiomelanocortin	Hypothalamic neuropeptide	96
PCSK1	Proprotein convertase 1	Processes pro-peptides (including POMC) to active moieties	62
MC4R	Melanocortin 4 receptor	Receptor for POMC products αMSH and βMSH	97,98
SIM1	SIM1, homologue of *Drosophila* single minded 1,	Transcription factor necessary for hypothalamic development	99

Disorders associated with developmental delay/cognitive dysfunction due to generalized disturbance of CNS function

Gene	Syndrome	Comment	Reference
–	Bardet–Biedl syndrome	At least 12 genes identified, most of which affect the function of the primary cilium	68
–	Prader–Willi syndrome	Imprinted locus on chromosome 15q including snoRNA cluster	100
GNAS	Pseudohypoparathyroidism	Encodes α-subunit of the stimulatory G protein	101
BDNF	WAGR syndrome	Deletion of *BDNF* in a subset of patients with WAGR syndrome is associated with obesity	102,103
NTRK2	–	Encodes TrkB, a receptor for brain-derived neurotrophic factor (BDNF) and neurotrophin 5 (NTF5). *NTRK2* mutations are associated with developmental delay and severe childhood obesity	104

Examples of human genes, mutations in which lead to a highly penetrant form of obesity. In disorders where careful phenotypic measurement has been undertaken, obesity is largely driven by increased appetite and/or diminished satiety. The mutations either have effects largely restricted to areas of the brain concerned with energy homeostasis, in which case obesity is usually the dominant presenting clinical feature, or impact on more generalized CNS functions, in which case obesity presents in the context of a more global developmental disorder. For a more comprehensive review of monogenic disorders leading to obesity see ref. 105.

were greatly facilitated by an explosion of research into the control of energy balance in obese animal models and in particular by the discovery of leptin as an adipocyte-derived hormone influencing the central control of energy balance[63], and the recognition of the brain melanocortin system as a major mediator of leptin action[64]. There are now at least 20 single gene disorders that clearly result in an autosomal form of human obesity (Table 3). Notably, so far all these disorders affect the central sensing and control of energy balance[65]. When energy balance is studied in detail in affected subjects it is clear that there is a major increase in appetite and reduction in satiety and that spontaneous measured food intake is greater than can be explained by increased body size[65]. In contrast, studies of energy expenditure in these subjects tend to reveal subtle if any decrement, although in MC4R deficiency there is a modest, but significant, tendency towards reduced metabolic rate[66] and reduced lipid oxidation[67]. The discovery that all of the genetic defects leading to Bardet–Biedl syndrome have a role in the structure and function of the primary cilium has focused attention on that organelle as a potentially crucial one in hypothalamic neurons responsible for the control of energy balance[68]. The marked reversal of the severe obesity seen in human leptin deficiency on administration of recombinant human leptin demonstrated the principle that, if a clear molecular basis for an individual's obesity could be found, then mechanism-based therapeutics could be highly effective[6].

Lessons from studies of the genetic basis for common obesity. Genome-wide association studies are beginning to identify the common genetic variation that underpins difference in adiposity across the normal population (Table 4). SNPs in the first intron of *FTO* were the first to emerge as unequivocally associated with human obesity[69,70]. *FTO* is a paradigm of how challenging it will be to move from proven genetic association to an understanding of the biology underpinning such an association. *FTO* is highly expressed in hypothalamus, where its expression is regulated by feeding and fasting[71]. Carriers of obesity-risk SNPs consistently show an increased appetite or measured food intake[72], and thus it seems clear that, like the monogenic disorders, the mechanism underlying the impact of this common genetic variant on human adiposity is principally through energy intake. However, there are still many questions. Intriguingly, mice rendered null for *Fto* are very small and have increased energy expenditure[73]. As yet, no one has demonstrated that the risk SNPs affect *FTO* expression. Although FTO has been shown to be a dioxygenase with an ability to demethylate 3-methylthymine in DNA *in vitro*, we don't know the true physiological substrate and we don't yet understand how its enzymatic function is linked to its role in energy balance[71].

The second obesity-risk SNP to be reported lies on chromosome 18, with its closest gene being *MC4R* (ref. 74). The association of this SNP with height[74] and with increased food intake[75] is reminiscent of the phenotype of severe MC4R deficiency and suggests that the SNP may indeed be operating through an effect on *MC4R*. Like *PPARG* and *KCNJ11* in diabetes, *MC4R* is an example of where disorders causing highly penetrant forms of a phenotype also harbour common variants which contribute to the phenotype in the broader population. Recently, several further SNPs robustly associated with obesity have been reported. Intriguingly, many of these are closely located to genes that are known to be expressed in the central nervous system[76].

Clinical implications of genetic discoveries. Leptin is a life-saving therapy for a very rare group of severely obese patients with congenital leptin deficiency. MC4R deficiency is present in up to 5% of severely obese children and we have argued that the determination of *MC4R* sequence should be a routine part of the evaluation of any severely obese child[65]. Although at the moment such information does not lead to any obvious specific therapy, it is important for carers, patients and parents to know that there is a powerful underlying biological drive promoting weight gain and that any success that diet and exercise have in restricting weight gain should be celebrated as an achievement. The definition of a clear genetic basis for a child's severe obesity may also prevent inappropriate and potentially highly damaging actions taken by health and social care agencies, who, in cases of severe childhood obesity, occasionally suggest temporary removal of a child from the parental home. As yet we know of only a small fraction of the genetic variants that underpin variation in adiposity in the general population, and it is unlikely that those SNPs which we know thus far will, on their own, have clinical utility. However, they have pointed the way towards new pathways, the pharmacological manipulation of which may ultimately be therapeutically beneficial.

Human genetics informs pathophysiology of obesity. Although it is likely that genetic variants influencing factors such as basal metabolic rate, or the propensity to take exercise, will, in time, be found, when we look at the information gleaned from the past 15 years of molecular genetic activity we cannot avoid concluding that, as much as type 2 diabetes is clearly a disease in which pancreatic beta-cell dysfunction is a critical element, obesity is a condition in which inherent genetic predisposition is dominated by the brain. Perhaps we will gradually come to see obesity not so much as a metabolic disease (although it can have grave metabolic consequences) but more of a neurobehavioural disorder, albeit one highly susceptible to the environment[65]. Most of the monogenic causes of human obesity seem to operate through increasing the 'set point' at which body adipose stores stabilize in the individual. Individuals with mutations in leptin, the leptin receptor and *MC4R*, for example, become obese at a very young age and remain severely, but not necessarily increasingly, obese throughout their lives. Other individuals, included among which are some of the most massively obese, gradually and progressively become more severely obese over time. Could some of these individuals have a progressive neurological disorder whereby instead of the adipostatic system simply functioning at an abnormal set point, the neuronal machinery controlling energy balance itself gradually degenerates? There is a dearth of information about hypothalamic histomorphometry in the severely obese, and this is a field ripe for further study.

Table 4 | Common genetic variants associated with body mass index

Chromosome	Nearby genes	How identified	Reference
1	*NEGR1**	GWAS meta-analysis, GWAS	76, 106
1	*SEC16B, RASAL2*	GWAS	106
2	*TMEM18*	GWAS meta-analysis, GWAS	76, 106
3	*SFRS10 (TRA2B), ETV5, DGKG*	GWAS	106
4	*GNPDA2*	GWAS meta-analysis	76
5	*PCSK1**	Candidate gene	107
6	*NCR3, AIF1, BAT2*	GWAS	106
10	*PTER†*	GWAS	108
11	*MTCH2*	GWAS meta-analysis	76
11	*LGR4, LIN7C, BDNF*	GWAS	106
12	*BCDIN3D, FAIM2*	GWAS	106
16	*MAF†*	GWAS	108
16	*SH2B1*, ATP2A1*	GWAS meta-analysis, GWAS	76, 106
16	*FTO**	GWAS	69, 70
18	*NPC1†*	GWAS	108
18	*MC4R**	GWAS	74
19	*KCTD15, CHST8*	GWAS meta-analysis, GWAS	76, 106

GWAS, genome-wide association study.
* Several of the genes most closely adjacent to the associated polymorphisms (mostly SNPs but some indels (insertions and deletions)) are expressed particularly highly in the CNS.
† Reported to be associated with early onset and/or extreme obesity but, so far, not confirmed in common adult obesity.

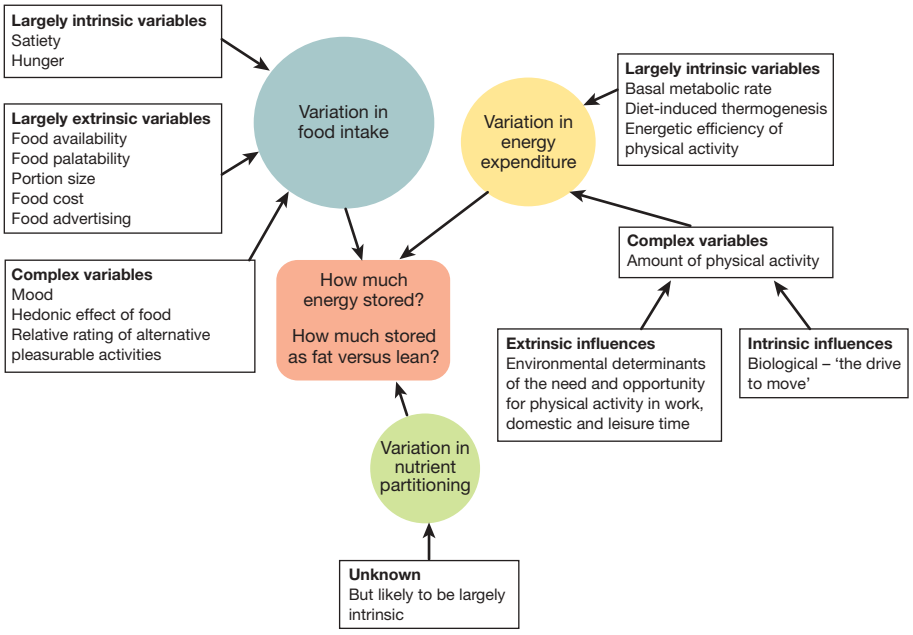
The remaining challenges

Although recent advances in the genetics of common metabolic disease have been exciting, the variants thus far detected only explain a very small fraction of the heritability of these disorders. It will be intriguing to see to what extent the remaining heritability is attributable to minor effects of common alleles, variation in gene copy number and rare mutations, and to watch the unfolding story of how they interact with environmental and epigenetic factors.

Box 2 | Variables influencing an individual's risk of becoming obese

Why are some people overweight and some people lean? This outcome results from the complex interaction between the cumulative intake and expenditure of energy and the tendency to deposit any excess of energy as either fat or lean mass (so-called nutrient partitioning). The major impact of molecular genetics on our understanding of the intrinsic variables that have an impact on energy balance is the unexpected finding that genetic variants causing severe familial obesity largely influence food intake through effects on hunger and satiety.

Importantly, many common SNPs broadly influencing adiposity across the human population are located in genes that are predominantly expressed in the brain. SNPs in the first intron of *FTO*, which are the most highly replicated common variants associated with human adiposity, are associated with alterations in appetite and food intake in humans.



Insulin resistance is a major feature of many common metabolic diseases and clearly has high heritability. It is perhaps surprising that so few genes having an impact on insulin action have thus far emerged from genome-wide approaches[77]. Gene discovery in this area is complicated by the major impact of recent diet and exercise on measures of insulin sensitivity and the fact that the measures of insulin sensitivity available in most sizeable epidemiological studies are indirect proxies that generally focus on the fasting state. The molecular mechanisms of insulin action in liver (the key target tissue in the fasting state) and both muscle and fat (the key target tissues in the post-prandial state) are known to have key differences, and it is likely that different sets of genetic variants will underpin inter-individual differences in basal and post-meal insulin sensitivity. The distribution of body fat has effects on insulin sensitivity and diabetes risk independent of its total amount and it is intriguing that SNPs affecting measures of fat distribution are now emerging[78].

As exemplified by *FTO* (see above), much effort and ingenuity will need to be expended to understand the precise mechanisms whereby risk alleles influence pathophysiology. This will need a multidisciplinary approach using cell biology, animal models and detailed pathophysiological studies in humans carrying risk variants.

Will the growing body of credible human genetic data help the pharmaceutical and biotechnology sectors in their efforts to design and develop better drug therapies for common metabolic diseases? In the 'drug discovery' part of the pipeline, knowledge that a particular potential therapeutic target, when genetically altered in humans, has an appropriate phenotype can provide reassurance about the relevance of that particular target. Additionally, knowledge that major loss- or gain-of-function mutations in that target do not result in other non-metabolic phenotypes in humans increases confidence that pharmacological manipulation of the target may have some specificity for the phenotype of interest. In this regard, information from rare highly penetrant mutations causing severe metabolic phenotypes may be of more direct utility to drug developers than common variants with subtle effects on metabolic phenotypes. Genetic studies of people with extreme familial leanness or severely obese subjects who avoid metabolic sequelae may be of particular interest.

Can human genetics accelerate or improve the development phase of the drug pipeline? When testing an antiobesity agent in terms of its ability to reduce metabolic risk it might, for example, be possible to enrich that group with genetic variants predisposing to type 2 diabetes so that an impact of weight loss on adverse metabolic outcome might be demonstrable earlier. The power of the currently available genetic information is probably insufficient for this to have a major impact at this stage. However, it would seem wise to continue to ensure that genetic material is obtained on participants from all clinical trials and to undertake post-hoc analysis examining whether one or other genetically defined subgroup had particularly pronounced therapeutic, or indeed adverse, effects.

Conclusions

Those of us who have grappled for decades with the challenge of understanding the fundamental mechanisms underlying type 2 diabetes and obesity are vulnerable to the assertion that we have been essentially misguided. It is clear that in environments where calories are difficult and expensive to access and much physical activity is required to be expended in acquiring them, obesity and type 2 diabetes are uncommon diseases[3]. It is difficult to refute the assertion that if modern populations returned to a hunter-gatherer state then obesity and diabetes would not be the major public heath threats that they now are. Nevertheless, the genetic loading that some unfortunate people receive is so adverse that they are likely to suffer metabolic disease despite their best efforts to avoid it. Human molecular genetics has allowed us to identify significant subsets of patients with obesity and/or diabetes where intrinsic biological factors have a major role, and some patients with these disease subtypes exhibit beneficial therapeutic responses to specific interventions targeted to underlying mechanism. It is also likely to have a continuing role in the validation of therapeutic targets for common forms of metabolic disease as well as the continued pathophysiological dissection of

superficially similar diseases into more refined and therapeutically relevant subtypes.

1. Maes, H. H., Neale, M. C. & Eaves, L. J. Genetic and environmental factors in relative body weight and human adiposity. *Behav. Genet.* **27,** 325–351 (1997).
2. Poulsen, P., Kyvik, K. O., Vaag, A. & Beck-Nielsen, H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance – a population-based twin study. *Diabetologia* **42,** 139–145 (1999).
3. Wild, S. *et al.* Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care* **27,** 1047–1053 (2004).
4. Wardle, J., Carnell, S., Haworth, C. M. & Plomin, R. Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. *Am. J. Clin. Nutr.* **87,** 398–404 (2008).
5. Pearson, E. R. *et al.* Switching from insulin to oral sulphonylureas in patients with diabetes due to Kir6.2 mutations. *N. Engl. J. Med.* **355,** 467–477 (2006).
Describes dramatic clinical benefits resulting from identification of a specific molecular subtype of diabetes.
6. Farooqi, I. S. *et al.* Beneficial effects of leptin on obesity, T cell hyporesponsiveness, and neuroendocrine/metabolic dysfunction of human congenital leptin deficiency. *J. Clin. Invest.* **110,** 1093–1103 (2002).
Describes reversal of severe human obesity in congenital leptin deficiency by administration of recombinant leptin.
7. World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF Consultation. ⟨http://www.idf.org/webdata/docs/WHO_IDF_definition_diagnosis_of_diabetes.pdf⟩ (2006).
8. Petersen, K. F. & Shulman, G. I. Etiology of insulin resistance. *Am. J. Med.* **119,** S10–S16 (2006).
9. Reaven, G. M. Banting lecture 1988. Role of insulin resistance in human disease. *Diabetes* **37,** 1595–1607 (1988).
10. Cline, G. W. *et al.* Impaired glucose transport as a cause of decreased insulin-stimulated muscle glycogen synthesis in type 2 diabetes. *N. Engl. J. Med.* **341,** 240–246 (1999).
11. Turner, R. C. *et al.* Pathogenesis of NIDDM—a disease of deficient insulin secretion. *Baillieres Clin. Endocrinol. Metab.* **2,** 327–342 (1988).
12. Tattersall, R. Maturity-onset diabetes of the young: a clinical history. *Diabet. Med.* **15,** 11–14 (1998).
13. Vionnet, N. *et al.* Nonsense mutation in the glucokinase gene causes early-onset non-insulin-dependent diabetes mellitus. *Nature* **356,** 721–722 (1992).
Describes first causative mutation to be identified in a familial form of diabetes.
14. Yamagata, K. *et al.* Mutations in the hepatocyte nuclear factor-4α gene in maturity-onset diabetes of the young (MODY1). *Nature* **384,** 458–460 (1996).
15. Yamagata, K. *et al.* Mutations in the hepatocyte nuclear factor-1α gene in maturity-onset diabetes of the young (MODY3). *Nature* **384,** 455–458 (1996).
Refs 14 and 15 highlight the importance of mutations in the HNF family of transcription factors in the causation of familial forms of diabetes.
16. Stoffers, D. A., Ferrer, J., Clarke, W. L. & Habener, J. F. Early-onset type-II diabetes mellitus (MODY4) linked to *IPF1*. *Nature Genet.* **17,** 138–139 (1997).
17. Horikawa, Y. *et al.* Mutation in hepatocyte nuclear factor-1 β gene (*TCF2*) associated with MODY. *Nature Genet.* **17,** 384–385 (1997).
18. Malecki, M. T. *et al.* Mutations in *NEUROD1* are associated with the development of type 2 diabetes mellitus. *Nature Genet.* **23,** 323–328 (1999).
19. Babenko, A. P. *et al.* Activating mutations in the *ABCC8* gene in neonatal diabetes mellitus. *N. Engl. J. Med.* **355,** 456–466 (2006).
20. Gloyn, A. L. *et al.* Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. *N. Engl. J. Med.* 350, 1838–1849 (2004); erratum **351,** 1470 (2004).
Refs 19 and 20 demonstrate that mutations constitutively activating the voltage-gated potassium channel in the pancreatic beta cell could lead to insulin-deficient diabetes from birth.
21. Støy, J. *et al.* Insulin gene mutations as a cause of permanent neonatal diabetes. *Proc. Natl Acad. Sci. USA* **104,** 15040–15044 (2007).
22. Gill-Carey, O. & Hattersley, A. T. Genetics and type 2 diabetes in youth. *Pediatr. Diabetes* **8** (Suppl. 9), 42–47 (2007).
23. Yoshimasa, Y. *et al.* Insulin-resistant diabetes due to a point mutation that prevents insulin proreceptor processing. *Science* **240,** 784–787 (1988).
24. Accili, D. *et al.* Insulin resistance due to mutations of the insulin receptor gene: an overview. *J. Endocrinol. Invest.* **11,** 857–864 (1992).
25. George, S. *et al.* A family with severe insulin resistance and diabetes due to a mutation in *AKT2*. *Science* **304,** 1325–1328 (2004).
26. Dash, S. *et al.* A truncation mutation in *TBC1D4* in a family with acanthosis nigricans and postprandial hyperinsulinemia. *Proc. Natl Acad. Sci. USA* **106,** 9350–9355 (2009).
27. McGarry, J. D. What if Minkowski had been ageusic? An alternative angle on diabetes. *Science* **258,** 766–770 (1992).
A highly influential commentary that helped switch attention from glucose to lipid metabolism in type 2 diabetes.
28. Barroso, I. *et al.* Dominant negative mutations in human PPARγ associated with severe insulin resistance, diabetes mellitus and hypertension. *Nature* **402,** 880–883 (1999).
The first evidence in any species that disruption of PPARγ function has a severe impact on insulin sensitivity.
29. Shackleton, S. *et al. LMNA*, encoding lamin A/C, is mutated in partial lipodystrophy. *Nature Genet.* **24,** 153–156 (2000).
30. Magré, J. *et al.* Identification of the gene altered in Berardinelli-Seip congenital lipodystrophy on chromosome 11q13. *Nature Genet.* **28,** 365–370 (2001).
31. Agarwal, A. K. *et al. AGPAT2* is mutated in congenital generalized lipodystrophy linked to chromosome 9q34. *Nature Genet.* **31,** 21–23 (2002).
32. Kim, C. A. *et al.* Association of a homozygous nonsense caveolin-1 mutation with Berardinelli-Seip congenital lipodystrophy. *J. Clin. Endocrinol. Metab.* **93,** 1129–1134 (2008).
33. Rubio-Cabezas, O. *et al.* Partial lipodystrophy and insulin resistant diabetes in a patient with a homozygous nonsense mutation in *CIDEC*. *EMBO Mol. Med* **1,** 280–287 (2009).
34. McCarthy, M. I. & Zeggini, E. Genome-wide association studies in type 2 diabetes. *Curr. Diab. Rep.* **9,** 164–171 (2009).
35. Stancáková, A. *et al.* Association of 18 confirmed susceptibility loci for type 2 diabetes with indices of insulin release, proinsulin conversion, and insulin sensitivity in 5,327 non-diabetic Finnish Men. *Diabetes* **58,** 2129–2136 (2009).
36. Pascoe, L. *et al.* Common variants of the novel type 2 diabetes genes *CDKAL1* and *HHEX/IDE* are associated with decreased pancreatic beta-cell function. *Diabetes* **56,** 3101–3104 (2007).
37. Grarup, N. *et al.* Studies of association of variants near the *HHEX, CDKN2A/B,* and *IGF2BP2* genes with type 2 diabetes and impaired insulin release in 10,705 Danish subjects: validation and extension of genome-wide association studies. *Diabetes* **56,** 3105–3111 (2007).
38. Palmer, N. D. *et al.* Association of *TCF7L2* gene polymorphisms with reduced acute insulin response in Hispanic Americans. *J. Clin. Endocrinol. Metab.* **93,** 304–309 (2008).
39. Palmer, N. D. *et al.* Quantitative trait analysis of type 2 diabetes susceptibility loci identified from whole genome association studies in the Insulin Resistance Atherosclerosis Family Study. *Diabetes* **57,** 1093–1100 (2008).
40. Bouatia-Naji, N. *et al.* A polymorphism within the *G6PC2* gene is associated with fasting plasma glucose levels. *Science* **320,** 1085–1088 (2008).
41. Bouatia-Naji, N. *et al.* A variant near *MTNR1B* is associated with increased fasting plasma glucose levels and type 2 diabetes risk. *Nature Genet.* **41,** 89–94 (2009).
42. Prokopenko, I. *et al.* Variants in *MTNR1B* influence fasting glucose levels. *Nature Genet.* **41,** 77–81 (2009).
43. Porte, D. Jr & Kahn, S. E. The key role of islet dysfunction in type II diabetes mellitus. *Clin. Invest. Med.* **18,** 247–254 (1995).
44. Tuomilehto, J. *et al.* Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N. Engl. J. Med.* **344,** 1343–1350 (2001).
45. Reaven, G. M. The fourth musketeer—from Alexandre Dumas to Claude Bernard. *Diabetologia* **38,** 3–13 (1995).
46. Højlund, K., Mogensen, M., Sahlin, K. & Beck-Nielsen, H. Mitochondrial dysfunction in type 2 diabetes and obesity. *Endocrinol. Metab. Clin. North Am.* **37,** 713–731 (2008).
47. Harding, H. P. & Ron, D. Endoplasmic reticulum stress and the development of diabetes: a review. *Diabetes* **51** (Suppl. 3), S455–S461 (2002).
48. Shoelson, S. E., Lee, J. & Goldfine, A. B. Inflammation and insulin resistance. *J. Clin. Invest.* **116,** 1793–1801 (2006).
49. van den Ouweland, J. M. *et al.* Mutation in mitochondrial tRNA(Leu)(UUR) gene in a large pedigree with maternally transmitted type II diabetes mellitus and deafness. *Nature Genet.* **1,** 368–371 (1992).
50. Inoue, H. *et al.* A gene encoding a transmembrane protein is mutated in patients with diabetes mellitus and optic atrophy (Wolfram syndrome). *Nature Genet.* **20,** 143–148 (1998).
51. Sandhu, M. S. *et al.* Common variants in *WFS1* confer risk of type 2 diabetes. *Nature Genet.* **39,** 951–953 (2007).
52. Fonseca, S. G. *et al.* WFS1 is a novel component of the unfolded protein response and maintains homeostasis of the endoplasmic reticulum in pancreatic beta-cells. *J. Biol. Chem.* **280,** 39609–39615 (2005).
53. Delépine, M. *et al. EIF2AK3*, encoding translation initiation factor 2-α kinase 3, is mutated in patients with Wolcott-Rallison syndrome. *Nature Genet.* **25,** 406–409 (2000).
54. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
55. Meigs, J. B. *et al.* Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N. Engl. J. Med.* **359,** 2208–2219 (2008).
56. Lyssenko, V. *et al.* Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N. Engl. J. Med.* **359,** 2220–2232 (2008).
57. Hattersley, A. T. & Pearson, E. R. Minireview: pharmacogenetics and beyond: the interaction of therapeutic response, beta-cell physiology, and genetics in diabetes. *Endocrinology* **147,** 2657–2663 (2006).
58. Kopelman, P. G. Obesity as a medical problem. *Nature* **404,** 635–643 (2000).
59. Jéquier, E. Energy expenditure in obesity. *Clin. Endocrinol. Metab.* **13,** 563–580 (1984).
60. Lissner, L. Measuring food intake in studies of obesity. *Public Health Nutr.* **5,** 889–892 (2002).
61. Montague, C. T. *et al.* Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature* **387,** 903–908 (1997).

62. Jackson, R. S. *et al.* Obesity and impaired prohormone processing associated with mutations in the human prohormone convertase 1 gene. *Nature Genet.* **16,** 303–306 (1997).
Refs 61 and 62 describe the first two genetic defects found to cause familial forms of human obesity.
63. Zhang, Y. *et al.* Positional cloning of the mouse obese gene and its human homologue. *Nature* **372,** 425–432 (1994).
Paper describing the discovery of leptin as an adipocyte-derived hormone with a role in the control of energy balance.
64. Huszar, D. *et al.* Targeted disruption of the melanocortin-4 receptor results in obesity in mice. *Cell* **88,** 131–141 (1997).
First definitive evidence for the role of melanocortin 4 receptor in the control of body weight.
65. O'Rahilly, S. & Farooqi, I. S. Human obesity: a heritable neurobehavioral disorder that is highly sensitive to environmental conditions. *Diabetes* **57,** 2905–2910 (2008).
66. Krakoff, J. *et al.* Lower metabolic rate in individuals heterozygous for either a frameshift or a functional missense *MC4R* variant. *Diabetes* **57,** 3267–3272 (2008).
67. Nogueiras, R. *et al.* The central melanocortin system directly controls peripheral lipid metabolism. *J. Clin. Invest.* **117,** 3475–3488 (2007).
68. Badano, J. L., Mitsuma, N., Beales, P. L. & Katsanis, N. The ciliopathies: an emerging class of human genetic disorders. *Annu. Rev. Genomics Hum. Genet.* **7,** 125–148 (2006).
69. Frayling, T. M. *et al.* A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316,** 889–894 (2007).
First description of a common human obesity gene discovered through a genome-wide association approach.
70. Dina, C. *et al.* Variation in *FTO* contributes to childhood obesity and severe adult obesity. *Nature Genet.* **39,** 724–726 (2007).
71. Gerken, T. *et al.* The obesity-associated *FTO* gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science* **318,** 1469–1472 (2007).
72. Cecil, J. E. *et al.* An obesity-associated *FTO* gene variant and increased energy intake in children. *N. Engl. J. Med.* **359,** 2558–2566 (2008).
73. Fischer, J. *et al.* Inactivation of the *Fto* gene protects from obesity. *Nature* **458,** 894–898 (2009).
74. Loos, R. J. *et al.* Common variants near *MC4R* are associated with fat mass, weight and risk of obesity. *Nature Genet.* **40,** 768–775 (2008).
This study reports a common variant in the vicinity of *MC4R* associated with human obesity.
75. Qi, L., Kraft, P., Hunter, D. J. & Hu, F. B. The common obesity variant near *MC4R* gene is associated with higher intakes of total energy and dietary fat, weight change and diabetes risk in women. *Hum. Mol. Genet.* **17,** 3502–3508 (2008).
76. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.* **41,** 25–34 (2009).
77. Rung, J. *et al.* Genetic variant near *IRS1* is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nature Genet.* **41,** 1110–1115 (2009).
78. Lindgren, C. M. *et al.* Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet.* **5,** e1000508 (2009).
79. Temple, I. K. *et al.* An imprinted gene(s) for diabetes? *Nature Genet.* **9,** 110–112 (1995).
80. Mackay, D. J. *et al.* Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in *ZFP57*. *Nature Genet.* **40,** 949–951 (2008).
81. Ræder, H. *et al.* Mutations in the *CEL* VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction. *Nature Genet.* **38,** 54–62 (2006).
82. Hearn, T. *et al.* Mutation of *ALMS1*, a large gene with a tandem repeat encoding 47 amino acids, causes Alström syndrome. *Nature Genet.* **31,** 79–83 (2002).
83. Collin, G. B. *et al.* Mutations in *ALMS1* cause obesity, type 2 diabetes and neurosensory degeneration in Alström syndrome. *Nature Genet.* **31,** 74–78 (2002).
84. McCarthy, M. I. & Hattersley, A. T. Learning from molecular genetics: novel insights arising from the definition of genes for monogenic and type 2 diabetes. *Diabetes* **57,** 2889–2898 (2008).
85. Zeggini, E. *et al.* Meta analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type-2 diabetes. *Nature Genet.* **40,** 638–645 (2008).
86. Deeb, S. S. *et al.* A Pro12Ala substitution in PPARγ2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. *Nature Genet.* **20,** 284–287 (1998).
The first association between a common SNP in PPARγ and type 2 diabetes.
87. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316,** 1331–1336 (2007).
88. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316,** 1336–1341 (2007).
89. Scott, L. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316,** 1341–1345 (2007).
90. Steinthorsdottir, V. *et al.* A variant in *CDKAL1* influences insulin response and risk of type 2 diabetes. *Nature Genet.* **39,** 770–775 (2007).
91. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445,** 881–885 (2007).
The first genome-wide association study in type 2 diabetes, focusing on more severe and early onset cases; identified SNPs are close to several islet genes.
92. Grant, S. F. *et al.* Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nature Genet.* **38,** 320–323 (2006).
The first and, so far, most significant common type 2 diabetes gene identified initially through linkage and positional candidate screening in a large Icelandic population.
93. Florez, J. C. Haplotype structure and genotype-phenotype correlations of the sulphonylurea receptor and the islet ATP-sensitive potassium channel gene region. *Diabetes* **53,** 1360–1368 (2004).
94. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in *TCF2* protects against type 2 diabetes. *Nature Genet.* **39,** 977–983 (2007).
95. Clément, K. *et al.* A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature* **392,** 398–401 (1998).
96. Krude, H. *et al.* Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by *POMC* mutations in humans. *Nature Genet.* **19,** 155–157 (1998).
97. Vaisse, C., Clement, K., Guy-Grand, B. & Froguel, P. A frameshift mutation in human *MC4R* is associated with a dominant form of obesity. *Nature Genet.* **20,** 113–114 (1998).
98. Yeo, G. S. *et al.* A frameshift mutation in *MC4R* associated with dominantly inherited human obesity. *Nature Genet.* **20,** 111–112 (1998).
Refs 97 and 98 are the first descriptions of human melanocortin 4 receptor deficiency, the commonest single gene disorder causing obesity in humans.
99. Holder, J. L. Jr, Butte, N. F. & Zinn, A. R. Profound obesity associated with a balanced translocation that disrupts the *SIM1* gene. *Hum. Mol. Genet.* **9,** 101–108 (2000).
100. Sahoo, T. *et al.* Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. *Nature Genet.* **40,** 719–721 (2008).
101. Spiegel, A. Albright's hereditary osteodystrophy and defective G proteins. *N. Engl. J. Med.* **322,** 1461–1462 (1990).
102. Gray, J. *et al.* Hyperphagia, severe obesity, impaired cognitive function, and hyperactivity associated with functional loss of one copy of the brain-derived neurotrophic factor (*BDNF*) gene. *Diabetes* **55,** 3366–3371 (2006).
103. Han, J. C. *et al.* Brain-derived neurotrophic factor and obesity in the WAGR syndrome. *N. Engl. J. Med.* 359, 918–927 (2008); erratum **359,** 1414 (2008).
104. Yeo, G. S. *et al.* A *de novo* mutation affecting human TrkB associated with severe obesity and developmental delay. *Nature Neurosci.* **7,** 1187–1189 (2004).
105. Beales, P. R., Farooqi, I. S. & O'Rahilly, S. (eds) *Genetics of Obesity Syndromes* (Oxford Univ. Press, 2009).
106. Thorleifsson, G. *et al.* Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature Genet.* **41,** 18–24 (2009).
107. Benzinou, M. *et al.* Common nonsynonymous variants in *PCSK1* confer risk of obesity. *Nature Genet.* **40,** 943–945 (2008).
108. Meyre, D. *et al.* Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature Genet.* **41,** 157–159 (2009).

Acknowledgements I thank M. Adams for assistance with this manuscript. I also thank A. Hattersley for many helpful suggestions, and I. Barroso, T. Coll, S. Farooqi, R. Loos, J. Rochford, D. Savage, R. Semple, N. Wareham and G. Yeo for discussions over many years. I acknowledge the support of the MRC Centre for Obesity and Related Metabolic Diseases, the Wellcome Trust, the NIHR Cambridge Biomedical Research Centre, and the EU for their continuing support of our work in this area.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence should be addressed to S.O.R. (so104@medschl.cam.ac.uk).

ARTICLES

Human DNA methylomes at base resolution show widespread epigenomic differences

Ryan Lister[1]*, Mattia Pelizzola[1]*, Robert H. Dowen[1], R. David Hawkins[2], Gary Hon[2], Julian Tonti-Filippini[4], Joseph R. Nery[1], Leonard Lee[2], Zhen Ye[2], Que-Minh Ngo[2], Lee Edsall[2], Jessica Antosiewicz-Bourget[5,6], Ron Stewart[5,6], Victor Ruotti[5,6], A. Harvey Millar[4], James A. Thomson[5,6,7,8], Bing Ren[2,3] & Joseph R. Ecker[1]

DNA cytosine methylation is a central epigenetic modification that has essential roles in cellular processes including genome regulation, development and disease. Here we present the first genome-wide, single-base-resolution maps of methylated cytosines in a mammalian genome, from both human embryonic stem cells and fetal fibroblasts, along with comparative analysis of messenger RNA and small RNA components of the transcriptome, several histone modifications, and sites of DNA–protein interaction for several key regulatory factors. Widespread differences were identified in the composition and patterning of cytosine methylation between the two genomes. Nearly one-quarter of all methylation identified in embryonic stem cells was in a non-CG context, suggesting that embryonic stem cells may use different methylation mechanisms to affect gene regulation. Methylation in non-CG contexts showed enrichment in gene bodies and depletion in protein binding sites and enhancers. Non-CG methylation disappeared upon induced differentiation of the embryonic stem cells, and was restored in induced pluripotent stem cells. We identified hundreds of differentially methylated regions proximal to genes involved in pluripotency and differentiation, and widespread reduced methylation levels in fibroblasts associated with lower transcriptional activity. These reference epigenomes provide a foundation for future studies exploring this key epigenetic modification in human disease and development.

Thirty-four years have passed since it was proposed that cytosine DNA methylation in eukaryotes could act as a stably inherited modification affecting gene regulation and cellular differentiation[1,2]. Since then, intense research effort has expanded our understanding of diverse aspects of DNA methylation in higher eukaryotic organisms. These include elucidation of molecular pathways required for establishing and maintaining DNA methylation, cell-type-specific variation in methylation patterns, and the involvement of methylation in multifarious cellular processes such as gene regulation, DNA–protein interactions, cellular differentiation, suppression of transposable elements, embryogenesis, X-inactivation, genomic imprinting and tumorigenesis[3–9]. DNA methylation, together with covalent modification of histones, is thought to alter chromatin density and accessibility of the DNA to cellular machinery, thereby modulating the transcriptional potential of the underlying DNA sequence[10].

Genome-wide studies of mammalian DNA methylation have previously been conducted, however they have been limited by low resolution[11], sequence-specific bias, or complexity reduction approaches that analyse only a very small fraction of the genome[12–14]. To improve our understanding of the genome-wide patterns of DNA methylation we have generated single-base-resolution DNA methylation maps throughout the majority of the human genome in both embryonic stem cells and fibroblasts. Furthermore, we have profiled several important histone modifications, protein–DNA interaction sites of regulatory factors, and the mRNA and small RNA components of the transcriptome to better understand how changes in DNA methylation patterns and histone modifications may affect readout of the proximal genetic information.

Single-base-resolution maps of DNA methylation for two human cell lines

Single-base DNA methylomes of the flowering plant *Arabidopsis thaliana* were previously achieved using MethylC-Seq[15] or BS-Seq[16]. In this method, genomic DNA is treated with sodium bisulphite (BS) to convert cytosine, but not methylcytosine, to uracil, and subsequent high-throughput sequencing. We performed MethylC-Seq for two human cell lines, H1 human embryonic stem cells[17] and IMR90 fetal lung fibroblasts[18], generating 1.16 and 1.18 billion reads, respectively, that aligned uniquely to the human reference sequence (NCBI build 36/HG18). The total sequence yield was 87.5 and 91.0 gigabases (Gb), with an average read depth of 14.2× and 14.8× per strand for H1 and IMR90, respectively (Supplementary Fig. 1a). In each cell type, over 86% of both strands of the 3.08 Gb human reference sequence are covered by at least one sequence read (Supplementary Fig. 1b), accounting for 94% of the cytosines in the genome.

We detected approximately 62 million and 45 million methylcytosines in H1 and IMR90 cells, respectively (1% false discovery rate (FDR), see Supplementary Information and Fig. 1a), comprising

[1]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA. [2]Ludwig Institute for Cancer Research, [3]Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, California 92093, USA. [4]ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Crawley, Western Australia 6009, Australia. [5]Morgridge Institute for Research, Madison, Wisconsin 53707, USA. [6]Genome Center of Wisconsin, Madison, Wisconsin 53706, USA. [7]Wisconsin National Primate Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53715, USA. [8]Department of Anatomy, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA.
*These authors contributed equally to this work.

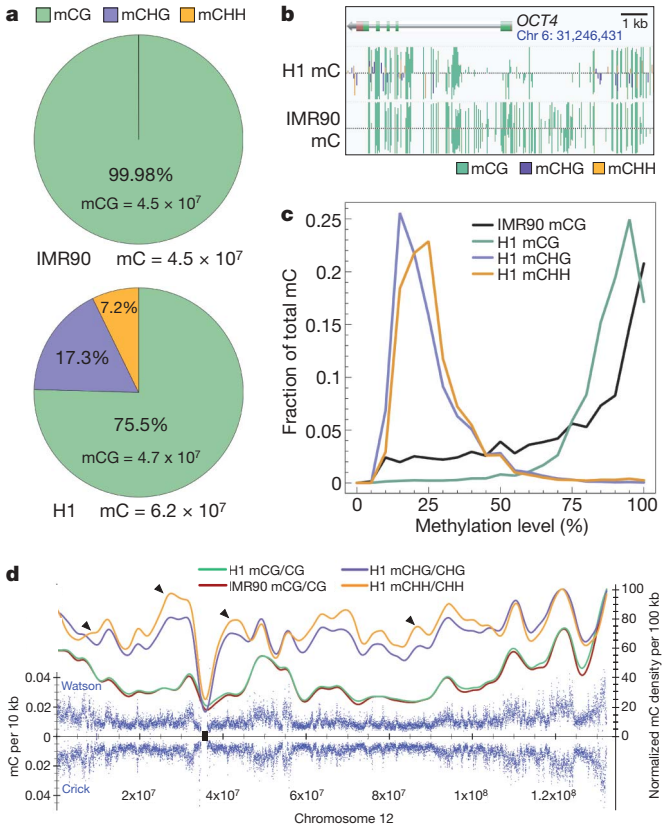


Figure 1 | Global trends of human DNA methylomes. **a**, The percentage of methylcytosines identified for H1 and IMR90 cells in each sequence context. **b**, AnnoJ browser representation of *OCT4*. **c**, Distribution of the methylation level in each sequence context. The *y* axis indicates the fraction of all methylcytosines that display each methylation level (*x* axis), where methylation level is the mC/C ratio at each reference cytosine (at least 10 reads required). **d**, Blue dots indicate methylcytosine density in H1 cells in 10-kb windows throughout chromosome 12 (black rectangle, centromere). Smoothed lines represent the methylcytosine density in each context in H1 and IMR90 cells. Black triangles indicate various regions of contrasting trends in CG and non-CG methylation. mC, methylcytosine.

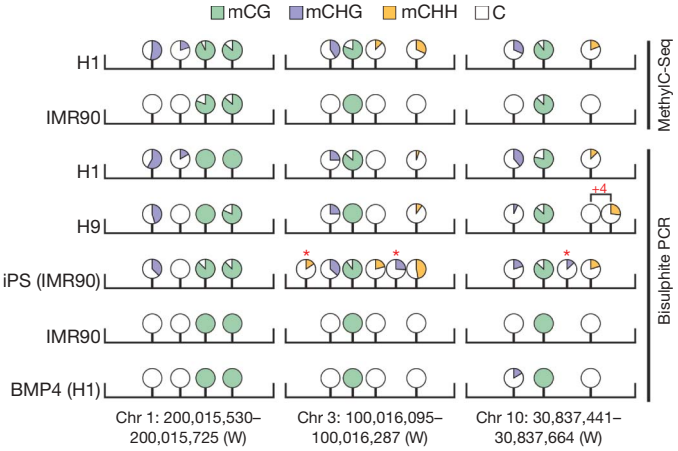


Figure 2 | Bisulphite-PCR validation of non-CG DNA methylation in differentiated and stem cells. DNA methylation sequence context is displayed according to the key and the percentage methylation at each position is represented by the fill of each circle (see Supplementary Table 2 for values). Non-CG methylated positions indicated by an asterisk are unique to that cell type and '+4' indicates a mCHH that is shifted 4 bases downstream in H9 cells. iPS, induced pluripotent stem cell.

5.83% and 4.25% of the cytosines with sequence coverage. Full browsing of the entire data set at single-base resolution can be performed at http://neomorph.salk.edu/human_methylome using the AnnoJ browser (http://www.annoj.org). Of the methylcytosines detected in the IMR90 genome, 99.98% were in the CG context, and the total number of mCG sites was very similar in both cell types. In the H1 stem cells we detected abundant DNA methylation in non-CG contexts (mCHG and mCHH, where H = A, C or T), comprising almost 25% of all cytosines at which DNA methylation is identified, and accounting for most of the difference in total methylcytosine number between the cell types (Fig. 1a). The prevailing assumption is that mammalian DNA methylation is located almost exclusively in the CG context; however, a handful of studies have previously detected non-CG methylation in human cells, and in particular in embryonic stem cells[19,20]. Bisulphite-PCR, cloning and sequencing of selected loci displaying H1 non-CG methylation in several human cell lines revealed that a second embryonic stem cell line, H9[17], displayed mCHG and mCHH at conserved positions, confirming that non-CG methylation is probably a general feature of human embryonic stem cells (Fig. 2, Supplementary Table 2). In addition, like IMR90 cells, BMP4-induced H1 cells lost non-CG methylation at several loci examined whereas methylation in the CG context was maintained, indicating that the pervasive non-CG methylation is lost upon differentiation. Furthermore, analysis of these loci in IMR90 induced pluripotent stem (iPS) cells revealed restored non-CG methylation (Fig. 2). Overall this demonstrates that the CHG and CHH methylation identified in H1 cells and absent in IMR90 cells is not simply due to genetic differences between the two cell types, but rather that the presence of non-CG methylation is characteristic of an embryonic stem-cell state. For each cell type, two biological replicates were performed with cells of different passage number (see Supplementary Information), and comparison of the methylcytosines identified independently in each replicate revealed a high concordance of cytosine methylation status between replicates (Supplementary Fig. 2). For each cell type, the final DNA methylation map presented in this study represents the composite of the two biological replicates. The *OCT4* gene (also called *POU5F1*) exemplifies both cell-specific differential methylation and the presence of non-CG methylation (Fig. 1b), and in addition displayed a ~50-fold reduction in *OCT4* transcript in IMR90 cells (data not shown). The absence of mCHG and mCHH methylation in IMR90 cells coincided with significantly lower transcript abundance of the *de novo* DNA methyltransferases (DNMTs) *DNMT3A* and *DNMT3B* and the associated *DNMT3L* in IMR90 cells (Supplementary Fig. 3), which is supported by a previous study of DNA methylation in embryonic stem cells and somatic cells[19] and by the determined target sequence specificity of these DNMTs[21,22].

Multiple reads covering each methylcytosine can be used as a readout of the fraction of the sequences within the sample that are methylated at that site[16], here referred to as the methylation level of a specific cytosine. Similar to the *Arabidopsis* genome[15], in the H1 genome we observed that 77% of mCG sites were 80–100% methylated, whereas 85% of mCHG and mCHH sites were 10–40% methylated (Fig. 1c), indicating that at sites of non-CG methylation only a fraction of the surveyed genomes in the sample was methylated. Notably, 56% of mCG sites in IMR90 cells were highly methylated (80–100%, Fig. 1c), indicating that although the total number of mCG sites in H1 and IMR90 cells is similar, in general the IMR90 mCG sites were typically less frequently methylated. In support of this, considering all CG site sequencing events, 82.7% and 67.7% were methylated in H1 and IMR90 cells, respectively. A global-scale view of DNA methylation levels revealed that the density of DNA methylation showed large variations throughout each chromosome (Fig. 1d). Sub-telomeric regions of the chromosomes frequently showed higher DNA methylation density (Fig. 1d and Supplementary Fig. 4), which was previously reported as being important for control of telomere length and recombination[23,24]. The smoothed profile of DNA methylation density in 100-kb windows indicated that on the chromosomal level the density profile of mCG in H1

and IMR90 cells was similar. The density profiles of mCHG and mCHH revealed that non-CG methylation was present throughout the entire chromosome. These two non-CG methylation marks showed a moderate correlation and did not always occur together (Pearson correlation 0.5 in 1-kb windows; Supplementary Fig. 2d). Notably, changes in density of the non-CG methylation were distinct from that of mCG in a number of regions.

Pervasive non-CG DNA methylation in embryonic stem cells

To characterize the abundant non-CG methylation in the H1 genome, we compared the average density of methylation relative to the underlying density of all potential sites of methylation in each context (henceforth referred to as the relative methylation density), throughout various genomic features (Fig. 3a and Supplementary Fig. 5). We observed a correlation in the density of mCG and the distance from the transcriptional start site (TSS), with mCG density increasing in the 5′ untranslated region (UTR) to a similar level in exons, introns and the 3′ UTR as to 2 kb upstream of the TSS (Fig. 3a). We generally observed lower relative densities of methylation at CG islands and TSS; however, a subset of these regions did not display this depletion (Supplementary Fig. 6)[13,14,25]. mCHG and mCHH methylation densities also decreased significantly towards the TSS and returned to the same level as 2 kb upstream at the end of the 5′ UTR; however, within exons, introns and 3′ UTRs the non-CG methylation densities were twice as high. Intriguingly, the mCHH density was approximately 15–20% higher in exons than within introns and the 3′ UTR. To identify links between gene activity and non-CG methylation level within the gene body we performed strand-specific RNA-Seq[15] and observed a positive correlation between gene expression and mCHG ($r = 0.60$) or mCHH ($r = 0.58$) density (Fig. 3b), with highly expressed genes containing threefold higher non-CG methylation density than non-expressed genes (Supplementary Fig. 7a). However, no correlation was observed

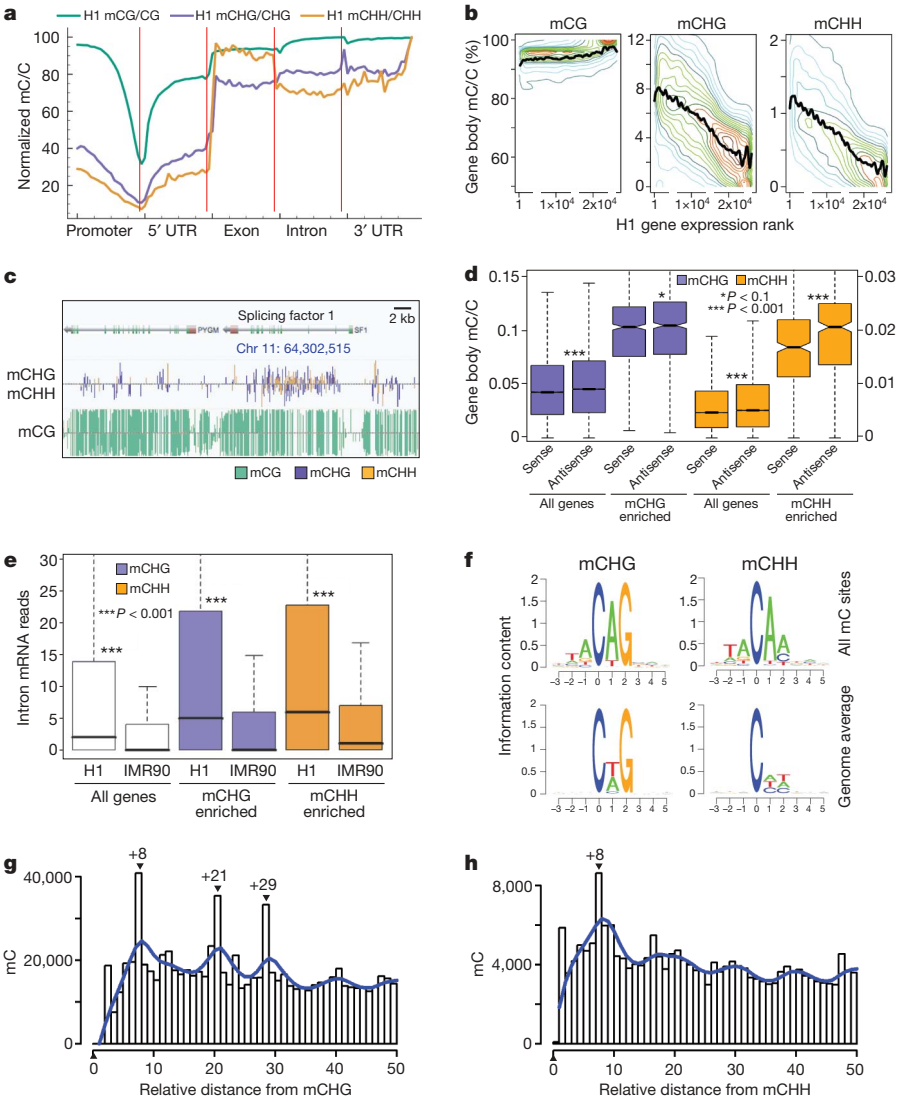


Figure 3 | Non-CG DNA methylation in H1 embryonic stem cells. a, Relative methylation density (the ratio of methylcytosines to reference cytosines) in H1 throughout different gene-associated regions (promoters encompass 2 kb upstream of the transcriptional start site).The mean mC/C profile was normalized to the maximum value. **b**, Relative methylation density within gene bodies (*y* axis) as a function of gene expression (*x* axis), with transcript abundance increasing from right to left. Coloured lines represent data point density and smoothing with cubic splines is displayed in black. **c**, Graphical representation of methylation at a non-CG methylation enriched gene, splicing factor 1. **d**, Average relative methylation densities in each sequence context within gene bodies on the sense or antisense strand relative to gene directionality. *P*-values for differences between sense and antisense densities are indicated. Boxes in **d** and **e** represent the quartiles and whiskers mark the minimum and maximum values. **e**, Number of mRNA intronic reads in all genes or genes associated with non-CG enriched regions, in H1 and IMR90 cells. *P*-values for differences between H1 and IMR90 reads are indicated. **f**, Logo plots of the sequences proximal to sites of non-CG DNA methylation in each sequence context in H1 cells. **g**, **h**, Prevalence of mCHG/mCHH sites (*y* axis) as a function of the number of bases between adjacent mCHG/mCHH sites (*x* axis) based on all non-redundant pairwise distances up to 50 nucleotides in all introns. Blue line represents smoothing with cubic splines.

between CG methylation density and gene expression in the H1 cells (Fig. 3b).

We identified 447 and 226 genes that were proximal to genomic regions highly enriched for mCHG and mCHH, respectively, with 180 genes in common. An example of non-CG methylation enrichment in such a gene, splicing factor 1 (*SF1*), is shown in Fig. 3c. Analysis of gene ontology terms for each set revealed significant enrichment for genes involved in RNA processing, RNA splicing and RNA metabolic processes ($P = 2 \times 10^{-11}$, Supplementary Fig. 7b). Unexpectedly, we found a significant enrichment of non-CG methylation on the antisense strand of gene bodies, for both mCHG and mCHH enriched sets of genes ($P < 0.1$ and $P < 0.001$, respectively, Fig. 3d). The antisense strand serves as the template for RNA polymerization, and further investigation will be required to determine if there are functional repercussions of this non-CG methylation strand bias. We also observed that genes in H1 had significantly more RNA originating from introns than in IMR90, relative to the total number of sequenced reads in each sample, and this discrepancy in intronic read abundance was significantly enhanced in the mCHG and mCHH enriched genes ($P < 0.001$, Fig. 3e). The higher abundance of intronic reads was associated with higher non-CG methylation within gene bodies, rather than differential non-CG methylation of exons versus introns.

In the *Arabidopsis* genome, the methylation state of a cytosine in the CG and CHG contexts is highly correlated with the methylation of the cytosine on the opposite strand within the symmetrical site[15,16]. Whereas we observed that 99% of mCG sites from the human cell lines were methylated on both strands, surprisingly mCHG was highly asymmetrical, with 98% of mCHG sites being methylated on only one strand. This raises an interesting question as to how these sites of DNA methylation are consistently methylated in a considerable fraction of the genomes without two hemi-methylated CHG sites as templates for faithful propagation of the methylation state (Fig. 1c). It is not yet known whether continual, but indiscriminate, *de novo* methyltransferase activity preferentially methylates particular CHG sites after replication, or if a persistent targeting signal is present that drives CHG methylation.

We analysed the genome sequence proximal to sites of non-CG methylation to determine whether enrichment of particular local sequences were evident, as previously reported in the *Arabidopsis* DNA methylomes[15,16]. Whereas no local sequence enrichment was observed for mCG sites, a preference for the TA dinucleotide upstream of non-CG methylation was observed (Fig. 3f and Supplementary Fig. 8). Furthermore, the base following a non-CG methylcytosine was most commonly an A, with a T also observed relatively frequently, a sequence preference observed in previous *in vitro* studies of the mammalian DNMT3 methyltransferases[21,22].

To determine whether there was any preference for the distance between adjacent sites of DNA methylation in the human genome, we analysed the relative distance between methylcytosines in each context within 50 nucleotides in introns. We focused on introns because these are genomic regions enriched in non-CG methylation, but unlike exons, are not constrained by protein coding selective pressures (Fig. 3g, h). Analyses for random genomic sequences and exons are presented in Supplementary Fig. 9, together with mCG spacing patterns. For methylcytosines in all contexts, a periodicity of 8–10 bases was evident (Fig. 3g, h and Supplementary Fig. 9), but interestingly a strong tendency was observed for two pairs of 8-base separated mCHG sites spaced with 13 bases between them. An 8–10 base periodicity was also evident for mCHH sites, corresponding to a single turn of the DNA helix, as previously observed in the *Arabidopsis* genome[16], indicating that the molecular mechanisms governing *de novo* methylation at CHH sites may be common between the plant and animal kingdoms. A structural study of the mammalian *de novo* methyltransferase DNMT3A and its partner protein DNMT3L found that two copies of each form a heterotetramer that contains two active sites separated by the length of 8–10 nucleotides in a DNA helix[26,27]. The consistent 8–10 nucleotide spacing we observed in the human genome suggests that DNMT3A may be responsible for catalysing the methylation at non-CG sites. Notably, the mCHG and mCHH relative spacing patterns were distinct, suggesting that this sub-categorization of the non-CG methylation is appropriate, and that distinct pathways may be responsible for the deposition of mCHG and mCHH in the human genome.

Depleted DNA methylation at DNA–protein interaction sites

Numerous past studies have documented that DNA methylation can alter the ability of some DNA binding proteins to interact with their target sequences[28–32]. To investigate this relationship further we used ChIP-Seq[33] to identify sites of protein–DNA interaction in H1 cells for a set of proteins important for gene expression in the pluripotent state, namely NANOG, SOX2, KLF4 and OCT4, as well as proteins involved in the transcription initiation complex and in enhancers (TAF1 and p300, respectively) (Supplementary Tables 3–8). In general we observed a decrease in the profile of relative methylation density towards the site of interaction, particularly in the non-CG context, independently from proximity to the TSS (Fig. 4a and Supplementary Fig. 10). The IMR90 genome showed higher average density of methylation at H1 SOX2 and p300 interaction sites, but had similar CG methylation densities for the H1 NANOG and OCT4 interaction sites,

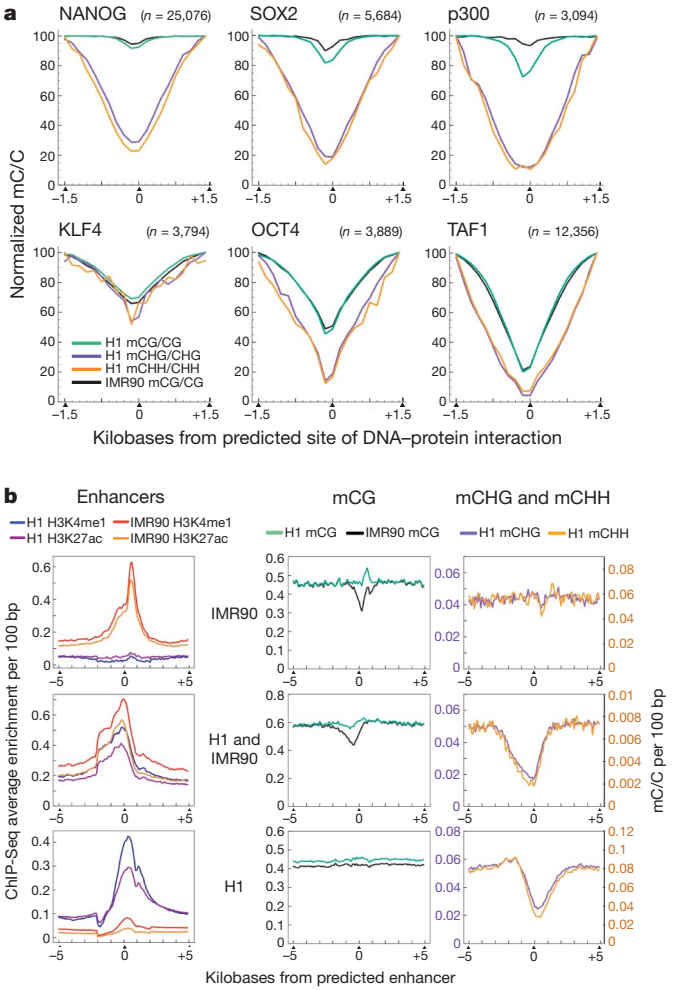


Figure 4 | Density of DNA methylation at sites of DNA–protein interaction. **a**, Average relative DNA methylation densities 1.5 kb upstream and downstream of predicted sites of DNA–protein interaction. **b**, Co-localization of H3K4me1 and H3K27ac ChIP-Seq tag enrichment indicative of enhancer sites that have been grouped into three sets: specific to IMR90 cells (top), H1 cells (bottom), or common to both H1 and IMR90 cells (middle). Average relative DNA methylation densities in each sequence context in 100-bp windows are displayed throughout 5 kb upstream and downstream of the enhancers in each of the sets.

even though the genes encoding these proteins are transcribed at a very low level in IMR90 relative to H1 cells (47–50-fold less mRNA), and are not considered to be functional in fibroblasts. This suggests that these genomic regions are generally maintained in a less methylated state in multiple cell types regardless of the occupancy of these specific DNA binding proteins.

We next analysed the patterns of DNA methylation in sets of enhancers either unique to each cell type or shared. ChIP-Seq was used to detect the location of enhancers throughout the H1 and IMR90 genomes, defined as regions of simultaneous enrichment of the histone modifications H3K4me1 and H3K27ac[34] (Fig. 4b). We examined the average relative DNA methylation density at enhancer sites, as well as the flanking genomic regions, and found a depletion of CG methylation at IMR90-specific enhancers, yet enrichment in mCG density in H1 at the same genomic locations (Fig. 4b). In contrast, at H1-specific enhancers there was no change in mCG density in either the H1 or IMR90 genome, but non-CG methylation density decreased approximately threefold at the enhancer sites, relative to the density 5 kb upstream and downstream. This is in agreement with the depletion of non-CG methylation in the H1 genome at predicted sites of p300 interaction (Fig. 4a), a strong indicator of enhancer activity[34]. The set of enhancer sites present in both H1 and IMR90 cells showed both of these cell-specific patterns: lower mCG density in IMR90 and lower non-CG methylation density in H1. The specific depletion of DNA methylation at active enhancers in each cell type (also recently reported on a limited basis[35]) indicates maintenance of these elements in an unmethylated state, potentially preventing interference in the process of protein–DNA interaction at these sites. Notably, H1 cells had depleted non-CG methylation but not mCG, in contrast to the mCG depletion at IMR90 enhancers. These data might indicate cell-type-specific utilization of different categories of DNA methylation, possibly coupled with novel stem-cell-specific factors that are able to recognize non-CG methylation, akin to the specific binding of the H3K9 histone methyltransferase KRYPTONITE to non-CG methylation sites in *Arabidopsis*[36].

Widespread cell-specific patterns of DNA methylation

The paradigm of DNA methylation controlling aspects of cellular differentiation necessitates that patterns of methylation vary in different cell types. Numerous studies have previously documented differences in DNA methylation between cell types and disease states[7,8,10,37]. With comprehensive maps of DNA methylation throughout the genomes of the two distinct cell types, we next characterized changes in DNA methylation evident between the H1 and

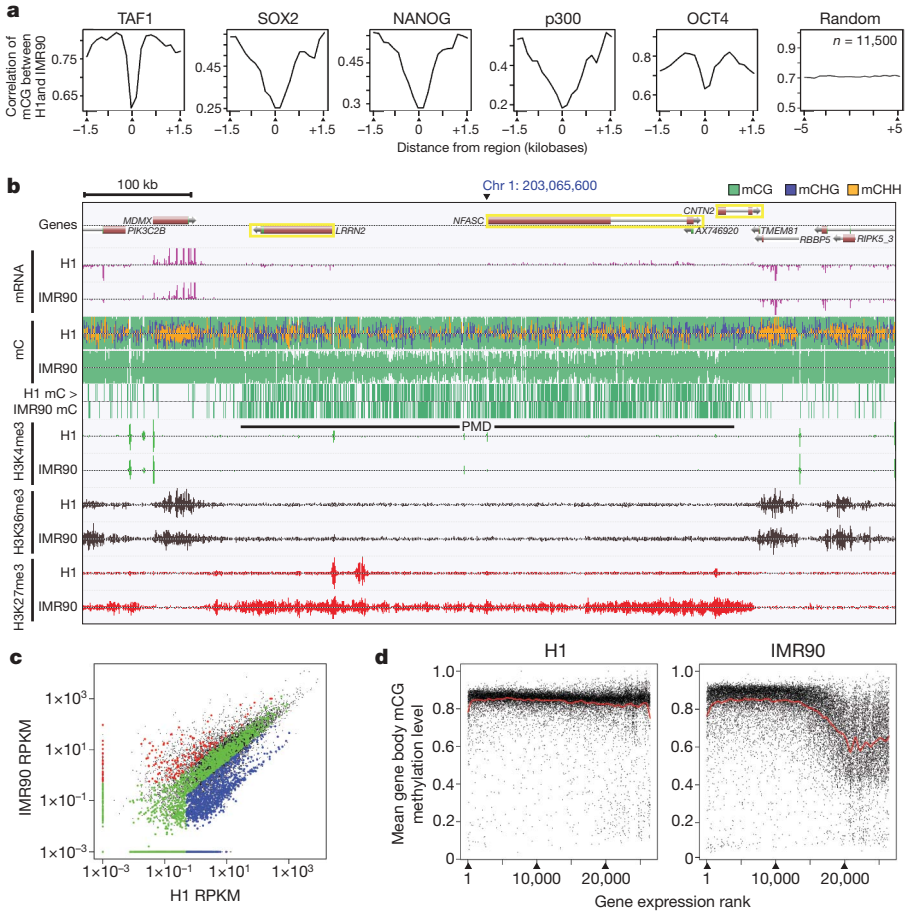


Figure 5 | Cell-type variation in DNA methylation. **a**, Pearson correlation coefficient of mCG methylation density (*y* axis) between H1 and IMR90 at various genomic features. Regions were divided into 20 equally sized bins from 5′ to 3′. Pearson correlation was determined in each bin considering all the H1 and IMR90 occurrences of the given genomic region. **b**, DNA methylation, mRNA and histone modifications in H1 and IMR90 cells associated with a PMD. Vertical lines above and below the dotted central line in DNA methylation tracks indicate methylcytosines on the Watson and Crick strands, respectively. Line vertical height indicates the methylation level. The H1 mC > IMR90 mC track indicates methylcytosines significantly more methylated in H1 than IMR90 at a 5% FDR (Fisher's exact test). Vertical bars in the mRNA and histone modification tracks represent sequence tag enrichment. A yellow box indicates any gene with ≥30-fold higher mRNA abundance in H1 than IMR90. **c**, Comparison of transcript abundance between H1 and IMR90 cells of genes with a transcriptional start site located in or within 10 kb of a PMD. Black dots indicate all genes in the genome; blue, red and green indicate PMD genes expressed ≥3-fold higher in H1, IMR90 or not differentially expressed, respectively. **d**, Mean gene body mCG methylation (at least 10 reads required) as a function of the gene expression rank, 1 being the most expressed. mC, methylcytosine; PMD, partially methylated domain; RPKM, reads per kilobase of transcript per million reads.

IMR90 DNA methylomes, and explored how these changes may relate to the distinctiveness of these cells.

Pairwise comparison of methylation at the same genomic coordinates between H1 and IMR90 is required to reveal cell-specific methylation patterns potentially masked by average profiles. The Pearson correlation coefficient of the mCG methylation state between H1 and IMR90 was calculated for 20 equally sized windows flanking or within various genomic features (Fig. 5a and Supplementary Fig. 11), providing a measure of methylation state conservation at these genomic features between the two cell types, and distinct from the average methylation density profiles presented above (Fig. 4). At the sites of protein–DNA interaction surveyed in Fig. 4a, we observed a decrease in the correlation of methylation compared to the flanking 1.5 kb of the genome (Fig. 5a), except for KLF4 (data not shown). This decrease was most pronounced at the predicted site of protein–DNA interaction, indicating that even though the mCG depletion was a general feature of the surveyed protein binding sites (Fig. 4a), when a pairwise comparison of the methylation status at each cytosine associated with the protein binding site between H1 and IMR90 was performed a significant decrease in the conservation of methylation was observed (Fig. 5a).

Surprisingly, we found that a large proportion of the IMR90 genome displayed lower levels of CG methylation than H1 (Fig. 1c). Contiguous regions with an average methylation level less than 70% were identified (mean length = 153 kb), which we termed partially methylated domains (PMDs) (Fig. 5b, Supplementary Fig. 12 and Supplementary Table 9). The PMDs comprised a large proportion of every autosome (average = 38.4%), and 80% of the IMR90 X chromosome (Supplementary Fig. 12), consistent with the lower levels of DNA methylation reported in the inactive X chromosome[38]. As IMR90 cells are derived from a female (XX), it is anticipated that simultaneous sequencing of BS-converted genomic DNA from both the inactive and the active X chromosomes will manifest as PMDs throughout the majority of the X chromosome. However, the widespread prevalence of PMDs on the autosomes was unexpected. We analysed the ratio of methylated to unmethylated CG sites within individual MethylC-Seq reads. The IMR90 reads located within PMDs were more frequently partially methylated or unmethylated compared to all IMR90 reads aligned to the autosomes (Supplementary Fig. 12b). The decrease in PMD methylation manifested similarly in IMR90 autosomes and chromosome X; however, currently we cannot determine whether common pathways are responsible for altering methylation patterns in all chromosomes.

Upon inspection of 5,644 genes with a TSS located in or within 10 kb of a PMD, we found a strong enrichment for these genes to be less expressed in IMR90 ($P = 2 \times 10^{-47}$, Fisher's exact test). Specifically, of all of the genes that were more highly expressed in H1 (H1 $\geq 3\times$ IMR90 transcript abundance), 42% were located within PMDs, compared to only 13% of all more highly expressed genes in IMR90 cells being located in PMDs (Fig. 5b, c and Supplementary Tables 10 and 11). Many of the partially methylated and downregulated genes in IMR90 displayed lower proximal H3K4me3 and H3K36me3 modifications, and higher proximal H3K27me3 levels (Fig. 5b, Supplementary Fig. 13 and R.D.H. *et al.*, submitted). Whereas in IMR90 cells we

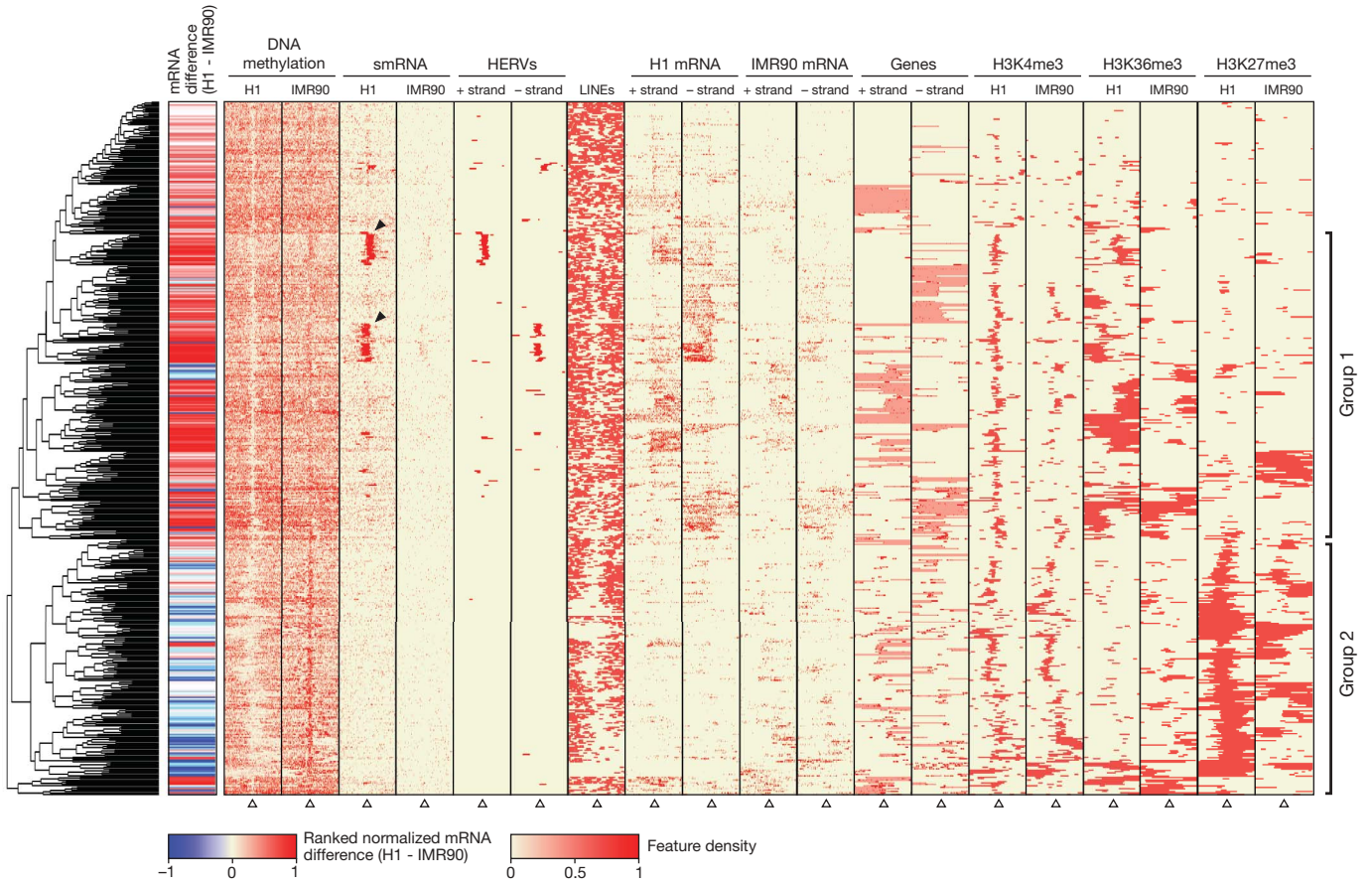


Figure 6 | Clustering of genomic, epigenetic and transcriptional features at differentially methylated regions. The density of DNA methylation, smRNA reads, strand-specific mRNA reads and the presence of domains of H3K4me3, H3k36me3 and H3K27me3 in H1 and IMR90 was profiled through 20 kb upstream and downstream of each of the 491 DMRs where DNA methylation was more prevalent in IMR90 than H1. Open triangles indicate the central point in each window. The side colour bar indicates the difference between H1 and IMR90 mRNA levels. The location of HERVs, LINEs and genes is displayed on each strand, where pink colouring indicates the gene body and dark red boxes represent exons. Black triangles indicate regions enriched for smRNAs that are coincident with HERVs. Group 1 and 2 are discussed in the text. DMRs, differentially methylated regions; HERVs, human endogenous retroviruses.

observed a positive correlation between the mean gene body mCG methylation level and gene expression, no such relationship was discernible in H1 cells (Fig. 5d). Consequently, the positive correlation between gene expression and gene body methylation recently reported[12] could be re-interpreted as a depletion of methylation in repressed genes in differentiated cells.

Stem cell hypomethylated regions

A sliding window approach was used to identify differentially methylated regions (DMRs) enriched for cytosines where IMR90 was more highly methylated than H1 (5% FDR, Fisher's exact test, Supplementary Fig. 14). We identified 491 DMRs, and in a window spanning 20 kb up- and downstream of each DMR we surveyed mCG density, mRNAs, small RNAs (smRNAs), H3K4me3, H3K36me3, H3K27me3, genes and repetitive elements (Fig. 6, Supplementary Table 12 and R.D.H. *et al.*, submitted). The DMRs were associated with 139 and 113 genes more highly expressed in H1 and IMR90, respectively. More than half of these genes were associated with DMRs located within 2 kb upstream of the TSS or the 5′ UTR, which include factors previously defined as having a role in embryonic stem-cell function[39] (Supplementary Fig. 15 and Supplementary Tables 13 and 14).

Complete linkage hierarchical clustering of these data revealed two broad categories of transcriptional activity, histone modifications and DNA methylation proximal to the DMRs (Fig. 6). Group 1 DMRs are associated with high proximal H3K4me3, H3K36me3 and transcriptional activity relative to IMR90, and are unmarked by H3K27me3 in both cell types. Although we did not observe widespread association of small RNA molecules with enrichment of DNA methylation, we found that a subset of group 1 DMRs co-localizes with dense clusters of small RNAs that map to annotated human endogenous retroviruses (HERVs)[40]. Notably, the HERVs were less densely methylated in H1 and frequently associated with high downstream transcriptional activity, in contrast to the more methylated state in IMR90 that was not associated with abundant small RNAs and showed little proximal transcription (Fig. 6 and Supplementary Fig. 16). Accurate targeting of DNA methylation by small RNAs is a well-established process in plants[41]. Although our data did not provide evidence for the existence of an analogous process in the human cells, further experiments may be required to investigate this relationship in greater detail, such as DNA methylation profiling following silencing of components of the RNA interference machinery.

Group 2 DMRs were associated with gene-rich sequences that were more highly expressed in IMR90 cells and generally exhibited a depletion of long interspersed nuclear elements (LINEs) in the flanking sequence, with concomitant H3K27me3 modification and less DNA methylation, as observed in many IMR90 PMDs. Furthermore, group 2 regions in H1 frequently displayed both H3K4me3 and H3K27me3 modifications, characteristic of the bivalent state that is thought to instil a suppressed but poised transcriptional status[42,43]. Many of these regions showed markedly less H3K27me3 in IMR90 cells in addition to more DNA methylation, suggesting that prior repression may have been relieved, and defining a set of genes potentially regulated by DNA methylation and involved in the developmental transition from a pluripotent to differentiated state.

Concluding remarks

We found extensive differences between the DNA methylomes of two human cell types, revealing the highly dynamic nature of this epigenetic modification. The genomic context of the DNA methylation is resolved, here revealing abundant methylation in the non-CG context, which is typically overlooked in alternative methodologies. Profiling of enhancers and different patterning of CG and non-CG methylation in gene bodies and their different correlation with gene expression suggest possible alternative roles for DNA methylation in these two contexts. The exclusivity of non-CG methylation in stem cells, probably maintained by continual *de novo* methyltransferase activity and not observed in differentiated cells, suggests that it may have a key role in the origin and maintenance of this pluripotent state. Essential future studies will need to explore the prevalence of non-CG methylation in diverse cell types, including variation throughout differentiation and its potential re-establishment in induced pluripotent states.

METHODS SUMMARY

Biological materials and sequencing libraries. Human H1, H9, BMP4-induced H1 and IMR90 cells were cultured as described previously[34,44,45]. smRNA-Seq libraries were generated from 10–50-nt small RNAs using the Small RNA Sample Prep v1.5 kit (Illumina), as per the manufacturer's instructions. Strand-specific mRNA-Seq libraries were produced using a modification of a protocol described previously[15]. Unique 5′ and 3′ RNA oligonucleotides were sequentially ligated to the ends of fragments of RNA isolated by depletion of rRNA from total RNA samples. MethylC-Seq libraries were generated by ligation of methylated sequencing adapters to fragmented genomic DNA followed by gel purification, sodium bisulphite conversion and four cycles of PCR amplification. ChIP-Seq libraries were prepared following Illumina protocols with minor modifications (See Supplementary Information). Sequencing was performed using the Illumina Genome Analyser II as per the manufacturer's instructions.

Read processing and alignment. MethylC-Seq sequencing data was processed using the Illumina analysis pipeline and FastQ format reads were aligned to the human reference genome (hg18) using the Bowtie alignment algorithm[46]. The base calls per reference position on each strand were used to identify methylated cytosines at 1% FDR. mRNA-Seq reads were aligned to the human reference and splice junctions of UCSC known genes using the ELAND algorithm (Illumina). smRNA-Seq reads that contained a subset of the 3′ adaptor sequence were selected and this adaptor sequence removed, retaining trimmed reads that were from 16 to 37 nucleotides in length. These processed reads were aligned to the human reference genome (NCBI build 36/HG18) using the Bowtie alignment algorithm, and any read that aligned with no mismatches and to no more than 1,000 locations in the reference genome was retained. Base calling and mapping of Chip-Seq reads were performed using the Illumina pipeline.

Received 19 June; accepted 21 September 2009.
Published online 14 October 2009.

1. Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187,** 226–232 (1975).
2. Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.* **14,** 9–25 (1975).
3. Bestor, T. H. The DNA methyltransferases of mammals. *Hum. Mol. Genet.* **9,** 2395–2402 (2000).
4. Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69,** 915–926 (1992).
5. Lippman, Z. *et al.* Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430,** 471–476 (2004).
6. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for *de novo* methylation and mammalian development. *Cell* **99,** 247–257 (1999).
7. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447,** 425–432 (2007).
8. Straussman, R. *et al.* Developmental programming of CpG island methylation profiles in the human genome. *Nature Struct. Mol. Biol.* **16,** 564–571 (2009).
9. Weber, M. & Schübeler, D. Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr. Opin. Cell Biol.* **19,** 273–280 (2007).
10. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Rev. Genet.* **10,** 295–304 (2009).
11. Rauch, T. A., Wu, X., Zhong, X., Riggs, A. D. & Pfeifer, G. P. A human B cell methylome at 100-base pair resolution. *Proc. Natl Acad. Sci. USA* **106,** 671–678 (2009).
12. Ball, M. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnol.* **27,** 361–368 (2009).
13. Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature Biotechnol.* **27,** 353–360 (2009).
14. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454,** 766–770 (2008).
15. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis. Cell* **133,** 523–536 (2008).
16. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452,** 215–219 (2008).
17. Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282,** 1145–1147 (1998).
18. Nichols, W. W. *et al.* Characterization of a new human diploid cell strain, IMR-90. *Science* **196,** 60–63 (1977).
19. Ramsahoye, B. H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl Acad. Sci. USA* **97,** 5237–5242 (2000).

20. Woodcock, D. M., Crowther, P. J. & Diver, W. P. The majority of methylated deoxycytidines in human DNA are not in the CpG dinucleotide. *Biochem. Biophys. Res. Commun.* **145,** 888–894 (1987).
21. Aoki, A. *et al.* Enzymatic properties of de novo-type mouse DNA (cytosine-5) methyltransferases. *Nucleic Acids Res.* **29,** 3506–3512 (2001).
22. Gowher, H. & Jeltsch, A. Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpA sites. *J. Mol. Biol.* **309,** 1201–1208 (2001).
23. Gonzalo, S. *et al.* DNA methyltransferases control telomere length and telomere recombination in mammalian cells. *Nature Cell Biol.* **8,** 416–424 (2006).
24. Steinert, S., Shay, J. W. & Wright, W. E. Modification of subtelomeric DNA. *Mol. Cell. Biol.* **24,** 4571–4580 (2004).
25. Brunner, A. L. *et al.* Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res.* **19,** 1044–1056 (2009).
26. Ferguson-Smith, A. C. & Greally, J. Epigenetics: perceptive enzymes. *Nature* **449,** 148–149 (2007).
27. Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A. & Cheng, X. Structure of Dnmt3a bound to Dnmt3L suggests a model for *de novo* DNA methylation. *Nature* **449,** 248–251 (2007).
28. Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* **405,** 482–485 (2000).
29. Clark, S. J., Harrison, J. & Molloy, P. L. Sp1 binding is inhibited by (m)Cp(m)CpG methylation. *Gene* **195,** 67–71 (1997).
30. Hark, A. T. *et al.* CTCF mediates methylation-sensitive enhancer-blocking activity at the *H19/Igf2* locus. *Nature* **405,** 486–489 (2000).
31. Kitazawa, S., Kitazawa, R. & Maeda, S. Transcriptional regulation of rat cyclin D1 gene by CpG methylation status in promoter region. *J. Biol. Chem.* **274,** 28787–28793 (1999).
32. Mancini, D. N., Singh, S. M., Archer, T. K. & Rodenhiser, D. I. Site-specific DNA methylation in the neurofibromatosis (NF1) promoter interferes with binding of CREB and SP1 transcription factors. *Oncogene* **18,** 4108–4119 (1999).
33. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316,** 1497–1502 (2007).
34. Heintzman, N. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).
35. Schmidl, C. *et al.* Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. *Genome Res.* **19,** 1165–1174 (2009).
36. Johnson, L. M. *et al.* The SRA methyl-cytosine-binding domain links DNA and histone methylation. *Curr. Biol.* **17,** 379–384 (2007).
37. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128,** 683–692 (2007).
38. Hellman, A. & Chess, A. Gene body-specific methylation on the active X chromosome. *Science* **315,** 1141–1143 (2007).
39. The International Stem Cell Initiative. Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. *Nature Biotechnol.* **25,** 803–816 (2007).
40. Villesen, P., Aagaard, L., Wiuf, C. & Pedersen, F. S. Identification of endogenous retroviral reading frames in the human genome. *Retrovirology* **1,** 32 (2004).
41. Chan, S. W. Inputs and outputs for chromatin-targeted RNAi. *Trends Plant Sci.* **13,** 383–389 (2008).
42. Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nature Cell Biol.* **8,** 532–538 (2006).
43. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).
44. Ludwig, T. *et al.* Feeder-independent culture of human embryonic stem cells. *Nature Methods* **3,** 637–646 (2006).
45. Ludwig, T. *et al.* Derivation of human embryonic stem cells in defined conditions. *Nature Biotechnol.* **24,** 185–187 (2006).
46. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Elwell and A. Hernandez for assistance with sequence library preparation and Illumina sequencing. R.L. is supported by a Human Frontier Science Program Long-term Fellowship. R.D.H. is supported by an American Cancer Society Postdoctoral Fellowship. This work was supported by grants from the following: Mary K. Chapman Foundation, The National Institutes of Health (U01 ES017166 and U01 1U01ES017166-01), the California Institute for Regenerative Medicine (RS1-00292-1), the Australian Research Council Centre of Excellence Program (CE0561495, DP0771156) and Morgridge Institute for Research, Madison, Wisconsin. We thank the NIH Roadmap Reference Epigenome Consortium (http://nihroadmap.nih.gov/epigenomics/referenceepigenomeconsortium.asp) and C. Gunter (Hudson-Alpha Institute) for assistance. This study was carried out as part of the NIH Roadmap Epigenomics Program.

Author Contributions Experiments were designed by J.R.E., B.R., R.L., J.A.T. and R.D.H. Cells were grown by J.A.-B. and Q.-M.N. MethylC-Seq, RNA-Seq and smRNA-Seq experiments were conducted by R.L. and J.R.N. ChIP-Seq experiments were conducted by R.D.H., L.L. and Z.Y. ChIP-Seq data analysis was performed by G.H., R.D.H. and L.E. BS-PCR validation was performed by R.H.D. Sequencing data processing was performed by R.L., J.T.-F., L.E., V.R. and G.H. Bioinformatic and statistical analyses were conducted by M.P., R.L., G.H., J.T.-F., R.H.D., R.S. and A.H.M. AnnoJ development was performed by J.T.F and A.H.M. The manuscript was prepared by R.L., M.P., R.H.D., A.H.M. and J.R.E.

Author Information Sequence data is available under the GEO accessions GSM429321-23, GSM432685-92, GSM438361-64, GSE17917, GSE18292 and GSE16256, and the SRA accessions SRX006782-89, SRX006239-41, SRX007165.1-68.1 and SRP000941. Analysed data sets can be obtained from http://neomorph.salk.edu/human_methylome. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.R.E. (ecker@salk.edu).

ARTICLES

RNA polymerase II–TFIIB structure and mechanism of transcription initiation

Dirk Kostrewa[1]*, Mirijam E. Zeller[2]*, Karim-Jean Armache[1]*†, Martin Seizl[1], Kristin Leike[1], Michael Thomm[2] & Patrick Cramer[1]

To initiate gene transcription, RNA polymerase II (Pol II) requires the transcription factor IIB (B). Here we present the crystal structure of the complete Pol II–B complex at 4.3 Å resolution, and complementary functional data. The results indicate the mechanism of transcription initiation, including the transition to RNA elongation. Promoter DNA is positioned over the Pol II active centre cleft with the 'B-core' domain that binds the wall at the end of the cleft. DNA is then opened with the help of the 'B-linker' that binds the Pol II rudder and clamp coiled-coil at the edge of the cleft. The DNA template strand slips into the cleft and is scanned for the transcription start site with the help of the 'B-reader' that approaches the active site. Synthesis of the RNA chain and rewinding of upstream DNA displace the B-reader and B-linker, respectively, to trigger B release and elongation complex formation.

Transcription of eukaryotic protein-coding genes begins with formation of the pre-initiation complex (PIC) on promoter DNA. The PIC contains Pol II and the general transcription factors, including B and the TATA-box-binding protein TBP[1]. Structural information exists for Pol II[2] and general factors[3], but the PIC structure remains controversial even at a topological level, and the structural transition from the PIC to the elongation complex is unclear. It was suggested that the PIC contains promoter DNA above the Pol II cleft (closed promoter complex), and that after DNA melting the template strand slips inside the cleft, resulting in the open promoter complex[4,5]. The RNA chain would then grow and form the eight-base-pair (bp) DNA–RNA hybrid observed in the elongation complex[6].

B has central roles in initiation. It recruits Pol II to the promoter, as its amino-terminal zinc ribbon domain ('B-ribbon') binds Pol II[7,8], and its carboxy-terminal domain ('B-core', comprising two cyclin folds) binds DNA and TBP[9] (Fig. 1a). B also functions after recruitment, because mutations in the region connecting the B-ribbon and the B-core compromise transcription and selection of the transcription start site (TSS), but not PIC formation[10–12]. B may be displaced owing to a clash with growing RNA[7,8]. The central roles of B are obvious in archaeal transcription, which requires only the B homologue transcription factor B (TFB), polymerase and TBP.

To understand the mechanism of initiation, the B polypeptide chain must be located on Pol II. Site-specific biochemical probing located the B-ribbon on the Pol II dock domain, the B-core on the wall, and the connecting region in the cleft[8,13]. X-ray analysis of the 10-subunit Pol II–B complex confirmed the position of the B-ribbon and revealed its orientation, but suggested that the B-core resides below the Pol II dock, and that the region connecting the B-ribbon and the B-core formed a hairpin (B-finger) that reached via the RNA exit tunnel near the active site, and returned through the tunnel to the dock[7]. Because of the different B-core locations, the two studies arrived at different PIC models. In one model, downstream promoter DNA runs over the cleft[13], whereas it runs over the outside of the clamp in the other[7].

Here we report the X-ray structure of B bound to the complete 12-subunit Pol II, which comprises the subcomplex Rpb4/7 that is required for initiation[14] but was lacking from the previous structure[7]. The structure locates B on Pol II, including the connecting region that consists of new 'B-reader' and 'B-linker' elements. It further results in a PIC model that resembles and refines the model derived biochemically[13], and allows modelling of the open complex. Mutational analysis in yeast and in a highly related archaeal system shows that the B-linker functions in promoter opening. Comparison with published data rationalizes TSS selection and B release after elongation-complex formation.

Pol II–B complex structure

The transient nature of the Pol II–B complex hampered co-crystallization of B with the complete yeast Pol II. Soaking of B into pre-formed Pol II crystals was also unsuccessful. We therefore prepared a recombinant Rpb4/7 variant in which the Rpb4 C terminus was fused to the N terminus of B by a linker of 20 glycine residues (Methods, for details see Supplementary Information). The linker length was more than twice the distance between the linked termini[7,15], and enabled the normal formation of a complex with TBP and synthetic nucleic acids. Samples of the complex, however, dissociated during crystallization and resulted in crystals of free Pol II–B complex. Diffraction beyond 7 Å resolution was very rarely observed, but complete diffraction data to 4.3 Å were eventually obtained.

Molecular replacement with the Pol II structure[15] resulted in difference electron density for B on the dock, in the RNA exit tunnel, in the cleft, and above the wall (Fig. 1 and Supplementary Information). We fitted the density on the dock with the free B-ribbon structure[16], aided by a peak for a zinc ion in the anomalous Fourier, and extended the B-ribbon at its C terminus through the RNA tunnel into a helical density underneath the lid. The densities above the wall were fitted with the N-terminal cyclin fold of the B-core[9]. No density was observed for the C-terminal cyclin fold, indicating mobility. Two remaining difference densities belonged to the B-linker in the cleft. An extended density formed a β-strand at the rudder (B-linker strand), whereas a helical density (B-linker helix) ran across the clamp coiled-coil above the rudder (Rpb1 helices α8 and α9).

[1]Gene Center Munich and Center for Integrated Protein Science Munich (CIPSM), Department of Chemistry and Biochemistry, Ludwig-Maximilians-Universität München, Feodor-Lynen-Strasse 25, 81377 Munich, Germany. [2]Institut für Biochemie, Genetik und Mikrobiologie, Universität Regensburg, Universitätsstrasse 31, 93053 Regensburg, Germany. †Present address: Department of Molecular Biology, Massachusetts General Hospital and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114, USA.
*These authors contributed equally to this work.

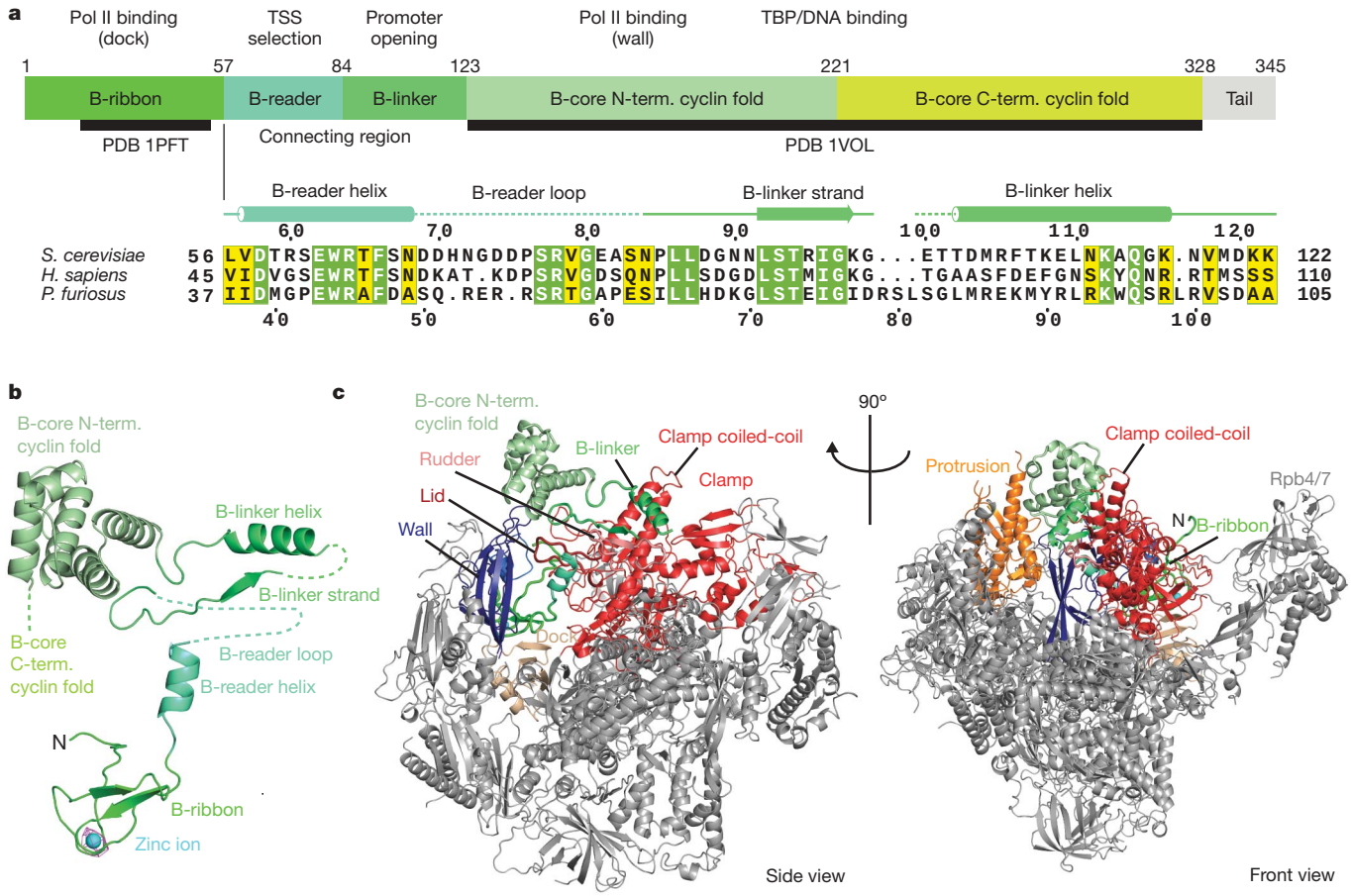


Figure 1 | Structure of Pol II–B complex. **a**, B domain organization and sequence conservation in the region connecting the B-ribbon and B-core. Yellow and green highlighting indicates conserved and invariant residues, respectively, between yeast (*S. cerevisiae*), human (*Homo sapiens*) and the archaeon *P. furiosus* (*Pfu*). **b**, Ribbon model of B as observed in its complex with Pol II. A peak in the anomalous difference Fourier (magenta) defines the zinc ion position (cyan sphere) in the B-ribbon. The view is from the side. **c**, Overview of the Pol II–B structure. Ribbon model with Pol II in silver and B in colours as in **a**. Side and front views are used. Pol II domains that interact with B are highlighted (dock, wheat; wall, blue; flap loop, light blue; clamp with coiled-coil, red; rudder, salmon; lid, dark red). The left view is from the side and lacks most of Rpb2, including the protrusion. The right view is from the front and includes Rpb2.

B binding induced small changes in the Pol II structure, including ordering of two loops on the wall that bind the B-ribbon (Rpb2 residues 879–883 and the Rpb2 flap loop 921–932). The structure was refined to an *R*-factor (R_{free}) of 22.0% (25.5%), and showed very good stereochemistry (Supplementary Information).

The structure (Fig. 1) confirmed the previously reported location of the B-ribbon[7,8], and the extension of the B-ribbon into the cleft towards the active site[7], but it deviates from the previous X-ray analysis[7] in three respects. First, most B residues previously assigned to a B-finger hairpin (residues 54–88; ref. 7) form a unidirectional extension of the B-ribbon into an α-helix and a mobile loop. We refer to this extension as B-reader, to indicate a proposed role in reading the DNA sequence during TSS selection (see later). Second, the B-core was observed above the wall, not at the dock. Third, the region between the B-reader and the B-core formed the B-linker in the cleft that was not observed previously. The discrepancies were resolved by calculating omit electron density maps with the deposited Pol II–B data[7]. These maps were consistent with the B-reader structure (Supplementary Information).

Taken together, the B polypeptide chain extends from the dock domain (B-ribbon) via the RNA exit tunnel to the hybrid-binding site and active centre (B-reader), to the rudder and clamp coiled-coil (B-linker), and to the Pol II wall (B-core) (Fig. 1c). These locations of B regions are apparently similar within the PIC, as they agree with biochemical mapping within the PIC[8,13] (Supplementary Information). B residues required for Pol II binding (Cys 45, Cys 48, Leu 50, Leu 52, Glu 62)[11,17] are near Pol II in the structure. B does not interact with Rpb4/7 in the structure, suggesting that the requirement of Rpb4/7 for initiation stems from a role in stabilizing the clamp, which binds the B-linker.

Models of closed and open complexes

We modelled a closed promoter complex by superimposing the B-core N-terminal cyclin folds in the Pol II–B and B-core–TBP–DNA complexes[9], and extending TATA DNA in both directions with B-DNA. In the model, TBP resides above the Pol II upstream face, downstream DNA runs between the clamp and protrusion along the edges of the cleft towards the jaws, and upstream DNA extends along the Rpb2 side (Fig. 2a). The model is similar to that obtained biochemically[13], except that the B-core is closer to the Pol II surface and rotated by approximately 60°, and the course of upstream DNA is altered accordingly.

To model the open complex, we assumed that the DNA template strand binds in the active centre as during elongation[18]. Consistently, placing the DNA from the elongation complex structure[19] into the Pol II–B structure did not lead to a clash between B and DNA. We further assumed that DNA opening commences 20 bp downstream of TATA[20]. The point of DNA opening lies above a tunnel lined by the B-core, B-linker, B-reader, and the Pol II protrusion, fork 1 and rudder (Fig. 2a). This 'template tunnel' leads to the upstream end of the DNA template strand in the elongation complex that could be connected to the point of DNA opening with only four nucleotides.

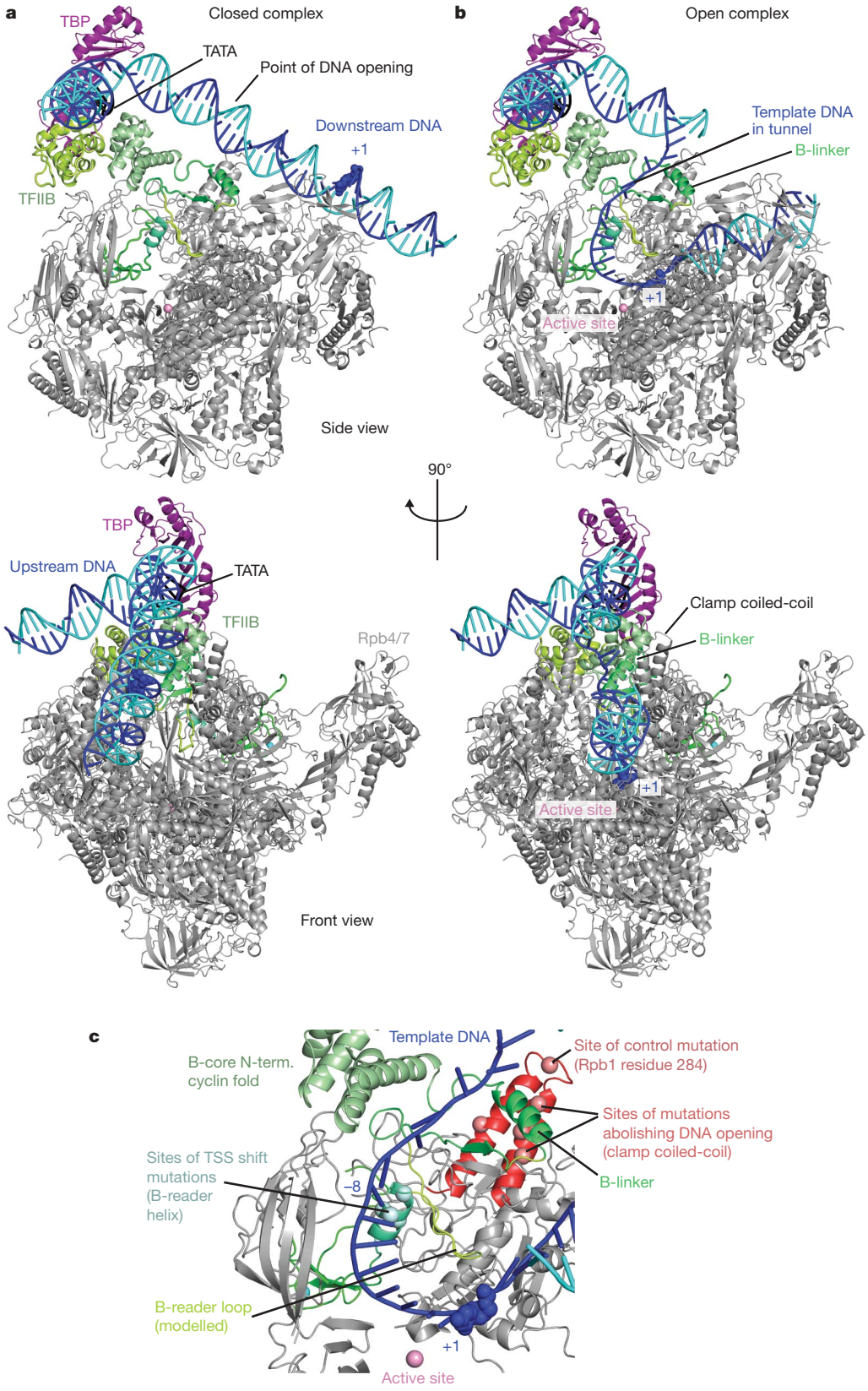


Figure 2 | Models of closed and open complexes. **a**, Model of the closed complex (minimal PIC). DNA template and non-template strands are in blue and cyan, respectively. The TATA element is in black and the nucleotide in the template strand that represents position +1 in the open complex is shown as a space-filling model. Top and bottom views are from the side and front, respectively. **b**, Model of the open complex. **c**, Location of nucleotides in DNA template strand initiator consensus sequence and mutations influencing start site selection and DNA opening. The open complex model is shown around the active centre. Positions −8 and +1 of the template strand are labelled. Position −8 lies adjacent to the B-reader helix that contains residues important for TSS selection (Glu 62, Trp 63, Arg 64, Phe 66, pale green spheres). The mobile B-reader loop (green-yellow), which contains residues Arg 78 and Val 79 required for initial transcription and TSS selection, could reach near positions −1 and +1. Sites of mutations abolishing DNA opening in archaeal transcription are shown as salmon spheres.

In the resulting open complex model (Fig. 2b), 34 nucleotides of DNA connect the upstream end of the TATA box to the TSS, which explains the minimal TATA–TSS distance in yeast[20]. The model is also consistent with slightly shorter TATA–TSS distances reported for other eukaryotic transcription systems, provided that DNA would be opened a few base pairs further upstream or the B-core would be slightly tilted on the wall. The model further explains TFB crosslinking to both DNA strands in the bubble of the related *Pyrococcus furiosus* (*Pfu*) open complex[21,22], and crosslinks to the template strand near the TSS[21]. The modelling suggests that the transition from the closed to the open complex involves a rotation of the downstream DNA duplex and slippage of the template strand into the template tunnel and active centre, probably aided by movement of the flexible fork loop 1 that is not required for transcription[18].

B-linker and DNA opening

The B-linker is located between closed DNA in the PIC model and the template single strand in the open complex model (Fig. 2), suggesting that it is involved in promoter opening. To test whether the B-linker is required for transcription *in vitro*, we used a nuclear extract from a temperature-sensitive yeast strain carrying a point mutation in the B gene *SUA7* (ref. 10). This extract enabled robust promoter-dependent transcription when recombinant wild-type B was added (Fig. 3a), but not when equally pure and stable B variants with B-linker mutations were added (Fig. 3b). Mutations in the linker strand also conferred lethality to yeast (Fig. 3c), showing that the B-linker is required for transcription *in vitro* and *in vivo*.

To test whether the B-linker is required for promoter opening, we used an archaeal transcription system based on *Pfu* polymerase, TFB and TBP[23]. Archaeal polymerase is highly homologous to Pol II[24,25], and forms a topologically similar PIC[21]. Four different *Pfu* TFB variants that altered the B-linker still formed PICs (Fig. 3d), but were inactive in promoter-dependent transcription (Fig. 3e), indicating a post-recruitment function for the B-linker. Mutation of four residues in the B-linker strand to alanines or destabilization of the B-linker helix by the point mutation Leu92Pro abolished transcription (Fig. 3e, g).

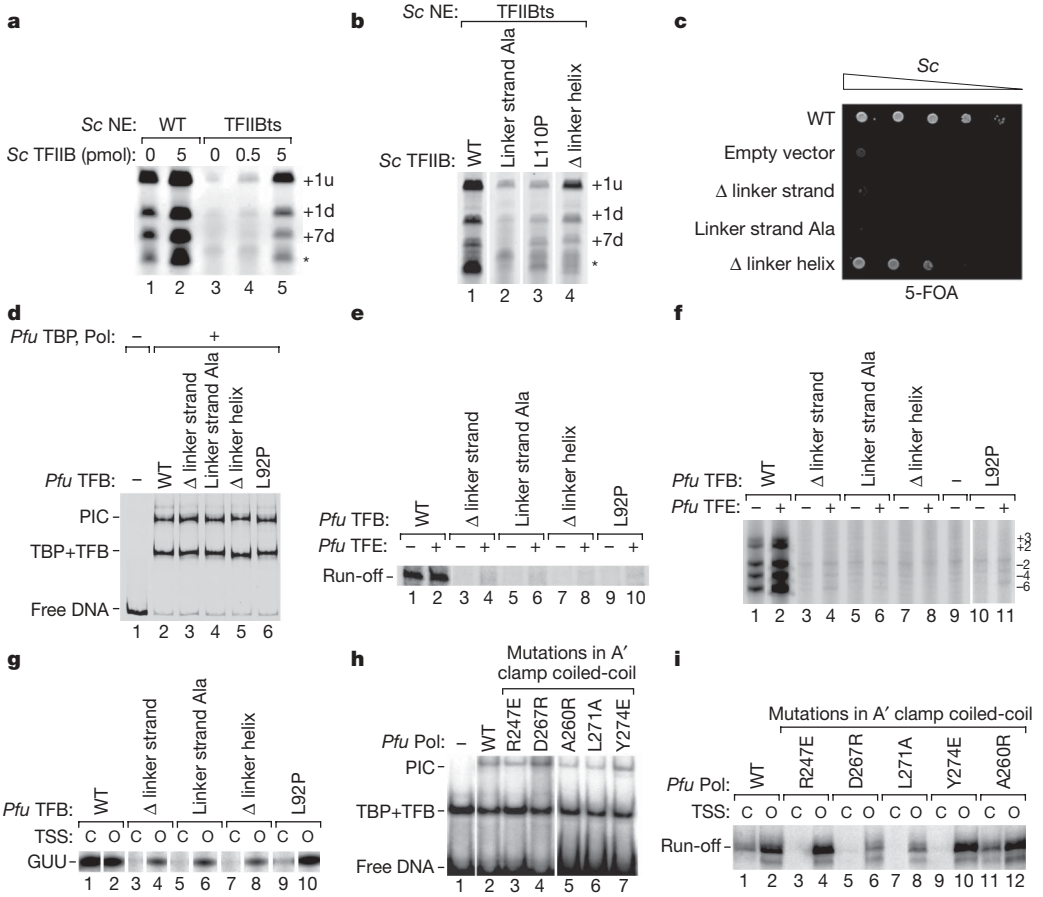


Figure 3 | B-linker and DNA opening. a, Nuclear extract (NE, 100 μg) from a temperature-sensitive yeast strain with a mutation in B (TFIIBts) is essentially inactive in *in vitro* transcription (lane 3). Activity is restored when recombinant wild-type (WT) B is added[10] (lane 5). Asterisk denotes a non-specific band. *Sc*, *S. cerevisiae*. The *HIS4-SNR14* template and the location of the observed TSSs (+1 upstream (u), +1 downstream (d), +7 downstream) are shown in Fig. 4a. **b**, B-linker mutations strongly affect *in vitro* transcription, but not TSS selection. The B variant 'Δ linker helix' lacks residues 104–114. The B variant 'linker strand Ala' corresponds to Leu92Ala/Ser93Ala/Thr94Ala/Ile96Ala. **c**, The B-linker strand is essential for yeast viability. A complementation assay using 5-fluoroorotic acid (5-FOA) selection is used. The variant 'Δ linker strand' lacks B residues 92–96. Other variants are as in **b**. **d**, B-linker variants of *Pfu* TFB form stable PICs. For the band-shift assay, a Cy5-labelled template containing the *gdh* promoter, TBP, TFB and RNA polymerase is used. The variants 'Δ linker strand' and 'Δ linker helix' lack TFB residues 71–75 and 85–95, respectively. The variant 'linker strand Ala' corresponds to Leu71Ala/Ser72Ala/Thr73Ala/Ile75Ala. **e**, B-linker mutations (compare with **d**) abolish run-off transcription by *Pfu* RNA polymerase. **f**, Permanganate footprinting shows that B-linker variants (compare with **d**) are defective in open complex formation. **g**, Mutations of the B-linker abolish transcription on a closed (C) template, but activity is rescued by a template with a pre-opened (O) transcription bubble containing mismatches at positions −1 to +2 relative to the TSS (+1). The band corresponding to the RNA product GUU is indicated. **h**, Reconstituted *Pfu* RNA polymerase variants form PICs according to band shift assays. The following point mutations were introduced in the B-linker-binding surface on the clamp coiled-coil of the polymerase subunit A′: Arg247Glu, Asp267Arg, Leu271Ala and Tyr274Ala (corresponding to Pol II Rpb1 residues Lys 271, Glu 291, Leu 295 and Phe 298, respectively). As a control, a conserved residue on the tip of the coiled-coil was mutated (Ala260Arg, corresponding to Pol II Rpb1 residue Ala 284). **i**, Mutations in the B-linker-binding surface on the clamp coiled-coil abolish transcription from a closed (C) DNA template but still support transcription from the pre-opened template (O, see **g**).

Mapping of single-stranded DNA regions by permanganate footprinting showed that the B-linker variants were unable to open the promoter, and the TFIIE homologue transcription factor E (TFE) could not rescue this defect (Fig. 3f). Transcription could, however, at least be partially rescued when we provided a RNA dinucleotide primer and a DNA template with a mismatched region that mimics part of the initial transcription bubble (Fig. 3g). Because free polymerase did not produce a transcript under the same conditions, the TFB variants with B-linker mutations were active except for DNA opening.

To test whether the polymerase surface on the clamp coiled-coil that binds the B-linker is required for transcription and promoter opening, we prepared four recombinant *Pfu* polymerase variants with different point mutations in the B-linker-binding surface (Fig. 2c). These variants still formed PICs to various extents (Fig. 3h), but were inactive in promoter-dependent transcription (Fig. 3i), whereas a control variant with a mutation outside the binding patch on top of the clamp coiled-coil was as active as wild-type polymerase (Figs 2c and 3i). However, all variants were able to transcribe a template with a mismatch bubble (Fig. 3i). Thus, the B-linker and its interaction surface on the clamp coiled-coil are required for transcription, but not for catalytic RNA synthesis, and, as demonstrated in the homologous archaeal system, this requirement stems from a role in promoter opening. This mechanism is apparently similar in Pol III initiation, which minimally requires TBP and the B-related factor Brf1, which contains an N-terminal region that is also required for promoter opening[26].

B-reader and DNA start site scanning

After open complex formation, yeast Pol II scans the DNA for an initiator (Inr) sequence motif that defines the TSS[20,27]. Several lines of evidence indicate that scanning involves threading of the template strand through the template tunnel until an Inr is detected by sequence-specific interactions. The template tunnel passes the active site, and Inr detection must occur near the active site where RNA synthesis is initiated. The template tunnel is flanked by the B-reader, and mutations in the B-reader residues Glu 62, Trp 63, Arg 64, Phe 66, Arg 78 and Val 79 shift the TSS, but do not influence PIC formation or promoter opening, as far as tested[11,12,17,28–30]. TSS shifts induced by B mutations depend on the Inr sequence[29]. Finally, Pol II and B are alone responsible for TSS selection[31].

To test the scanning model, we prepared a transcription template by fusing the *HIS4* promoter sequences around the TATA box to the duplicated Inr of the *SNR14* promoter[27] (Fig. 4a). In an *in vitro* assay[10], this fusion template generated transcripts that initiated at previously mapped TSSs in the *SNR14* promoter[27] (Fig. 4b, lane 2). Thus, the Inr alone determined the TSS, consistent with the scanning model. The yeast Inr consensus motif comprises a conserved A residue at position −8, and a CA or TG dinucleotide at positions −1/+1 of the non-template strand[32]. In the open complex model, the complementary template T residue at −8 is adjacent to the B-reader helix (Fig. 2c). Consistent with this modelling and published data[11,12,17,27–30], mutation of these B-reader residues decreased transcription efficiency and in some cases led to detectable changes in TSS selection (Fig. 4c). Mutation of the −8 position in the first Inr of the fusion promoter led to an almost complete loss of the corresponding transcript (Fig. 4b). When both positions −8 and −7 were mutated, transcription was totally lost (Fig. 4b). These results show that the −8 position of the Inr is important for TSS selection and suggest that it is recognized with the help of the B-reader helix.

The flexible B-reader loop may contribute to the recognition of the Inr dinucleotide at positions −1/+1, and to the stabilization of the first two nucleoside triphosphate (NTP) substrates in the active centre in the initiating complex. First, the B-reader loop contains the invariant residue Arg 78 that may reach near the active site (Fig. 2c). Arg 78 is not required for PIC formation, but its mutation leads to TSS shifts and altered abortive transcription *in vitro*, and causes sensitivity to NTP depletion or lethality *in vivo*[28,33]. Second, mutation of the neighbouring Val 79 also leads to TSS shifts, and most strongly effects TSS selection when the dinucleotide at positions −1/+1 is changed[29]. Third, NTP depletion can cause alternative TSS selection[34–36]. The B-reader loop may also stabilize the open bubble, because mutation of the residue corresponding to Arg 78 in *Pfu*

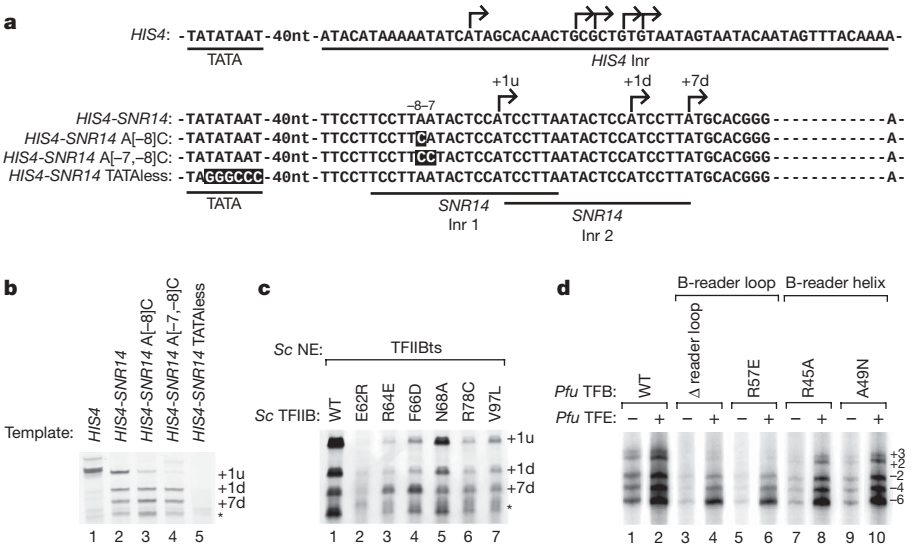


Figure 4 | B-reader and DNA start site scanning. **a**, Design of fusion promoter *HIS4-SNR14*. In promoters A[−8]C and A[−7,−8]C, adenines at position −8 or at −8 and −7, respectively, relative to the first TSS were replaced by cytosines (white). Arrows indicate *in vitro* TSSs in *HIS4* and *HIS4-SNR14* fusion promoters, respectively (compare with Fig. 3a, b). nt, nucleotides. **b**, Three different TSSs are used during *in vitro* transcription in yeast extracts with the *HIS4-SNR14* fusion promoter (lane 2). The mutations in promoters A[−8]C and A[−7,−8]C gradually eliminate recognition of the first TSS (lanes 3 and 4). **c**, Mutations in the yeast B-reader residues Arg 64, Phe 66, Arg 78 and Val 79, but not Asn 68, lead to TSS shifts. **d**, The *Pfu* TFB B-reader loop is required for opening and/or stabilization of the bubble just downstream of the TSS (position +3 and +2). Permanganate footprints are lost at downstream positions after deletion of part of the B-reader loop ('Δ reader loop', lacking residues 55–60) or mutation of residue Arg 57 in this loop (corresponding to Arg 78 in yeast). In contrast, mutations in the B-reader helix (Arg45Ala and Ala49Asn, corresponding to yeast residues Arg 64 and Asn 68, respectively) result in partial opening defects that are rescued by TFE.

(Arg 57) impairs DNA opening around the TSS *in vitro* whereas mutations in the B-reader helix do not (Fig. 4d). Consistently, defects arising from mutation of the B-reader are rescued by TFE (Fig. 4d), which stabilizes the bubble[37]. The flexibility of the B-reader loop may enable maintenance of nucleic acid contacts during initial growth of the RNA, to stabilize the early transcribing complex.

Initiation–elongation transition

Published data and our results converge on a model for transcription initiation, including the transition to elongation. First, promoter DNA is recruited to Pol II by the B-ribbon that binds the dock, and is positioned above the cleft by the B-core that binds the wall (closed complex formation). Second, DNA is melted around 20 bp downstream of the TATA box with the help of the B-linker, the emerging template strand slips in the cleft to fill the template tunnel, the bubble is stabilized near the active centre with the help of the B-reader, and the downstream DNA duplex is loaded into the downstream cleft (open complex formation). Third, the template strand is threaded through the template tunnel along the active site and is scanned for an Inr motif with the help of the B-reader (DNA start site scanning). Fourth, the first two NTPs are positioned opposite a conserved Inr dinucleotide motif and the first phosphodiester bond is formed (RNA chain initiation). Fifth, RNA growth results in short DNA–RNA hybrids that are transiently bound, leading to frequent ejection of short RNAs, maybe due to interference of the growing RNA with the B-reader loop (abortive transcription). Sixth, growth of the RNA beyond seven nucleotides triggers B release and formation of the elongation complex (promoter escape).

Comparison of our models with the complete elongation-complex model[38] shows that the initiation–elongation transition markedly alters the path of upstream DNA, which runs towards Rpb2 on the back of Pol II in the PIC (Fig. 5b), but emerges from the top of the cleft in the elongation complex. This comparison further suggests that B release during promoter escape results from a clash of RNA with the B-reader helix from position −8 onwards, and from a clash of the upstream DNA duplex and non-template DNA strand in the elongation complex with the B-linker above the rudder (Fig. 5a). Consistently, B release during promoter escape requires an RNA length of at least seven nucleotides and rewinding of upstream DNA[39].

The initiation–elongation transition also requires the general factors TFIIE, -F, and -H that we placed on the closed complex model (Fig. 5b). TFIIH is required for DNA opening, and apparently binds near the jaws at the entrance to the cleft because it crosslinks to downstream DNA[40]. TFIIE binds to the clamp[41] and its archaeal homologue TFE stabilizes the open complex[37], suggesting it traps the non-template strand in the initial bubble during open complex formation. TFIIF binds on the opposite side of the cleft[41] where it may form contacts to downstream and upstream DNA[40] (Fig. 5b).

Comparison to the bacterial system

Whereas Pol II is related to bacterial RNA polymerase in sequence and structure[42], the Pol II general factors show little or no sequence similarity to the bacterial σ factors. Comparison of our results with structures of bacterial RNA-polymerase–σ70 complexes in free form[43,44] and with upstream promoter DNA[45], however, shows topological similarities in the way B and σ70 bind to the polymerases. The B-ribbon, B-linker and B-core bind to polymerase surfaces that are also bound by parts of the σ70 domains σ_4, σ_2 and σ_3, respectively, and the B-reader corresponds topologically to the σ3.2 linker (Fig. 6a). The B-ribbon and σ_4 are both bound by a helix in the flap loop[46] (or flap tip, Fig. 6a) that is required for σ function[47]. The B-reader and σ3.2 both contain an extended, partially helical segment passing below the lid, and a mobile loop that extends near the active site and can be crosslinked to NTP in the case of the bacterial factor[48]. However, the B and σ polypeptide chains show opposite directionality and the domain structures differ (Fig. 6a), suggesting that B and σ result from convergent evolution.

The B-linker and σ_2 are both involved in promoter opening and both bind the clamp coiled-coil and rudder, which are also implicated

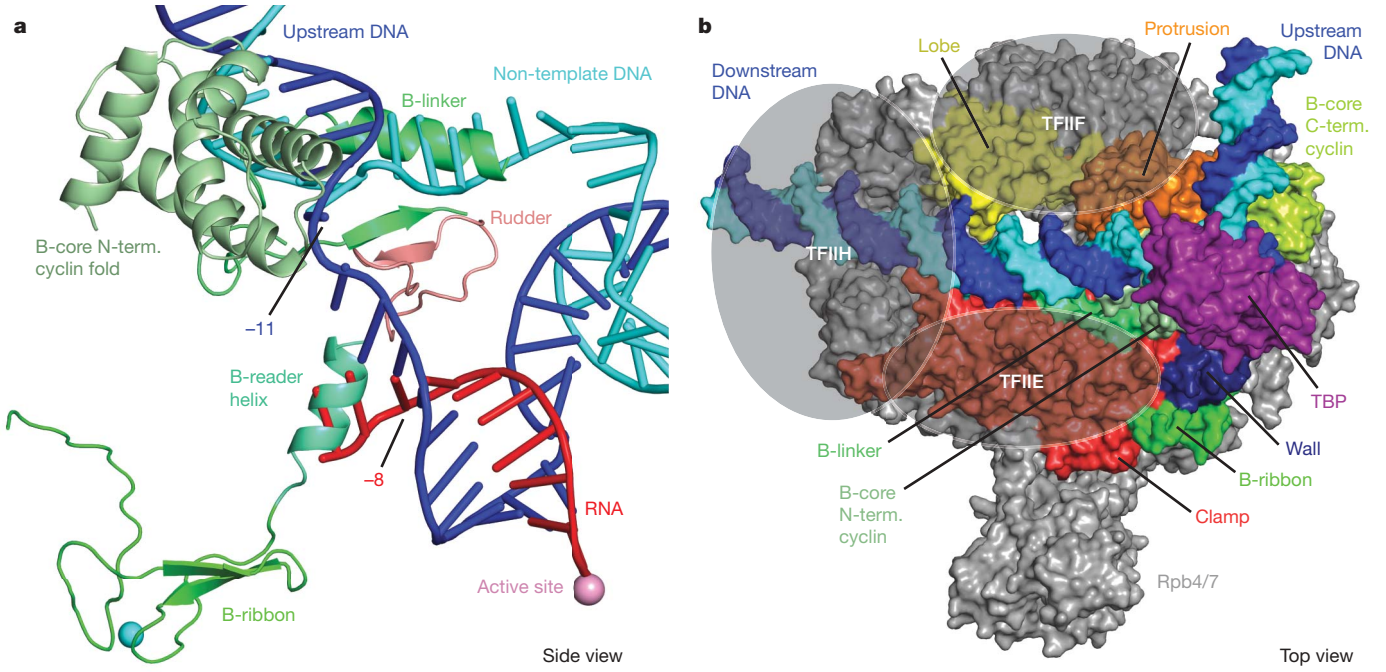


Figure 5 | Initiation–elongation transition and PIC architecture. a, Clash of growing RNA with the B-reader and clash of upstream DNA and non-template DNA with the B-linker. The Pol II–B complex structure was superimposed with the complete Pol II elongation complex structure[19] extended by upstream and non-template DNA[38]. The view is from the side. **b**, Architecture of the complete PIC. The closed complex model (Fig. 2a) is shown as a surface representation from the top. Locations of the general factors TFIIE, -F and -H, as determined by biochemical probing[40,41], are indicated by semi-transparent ellipsoids.

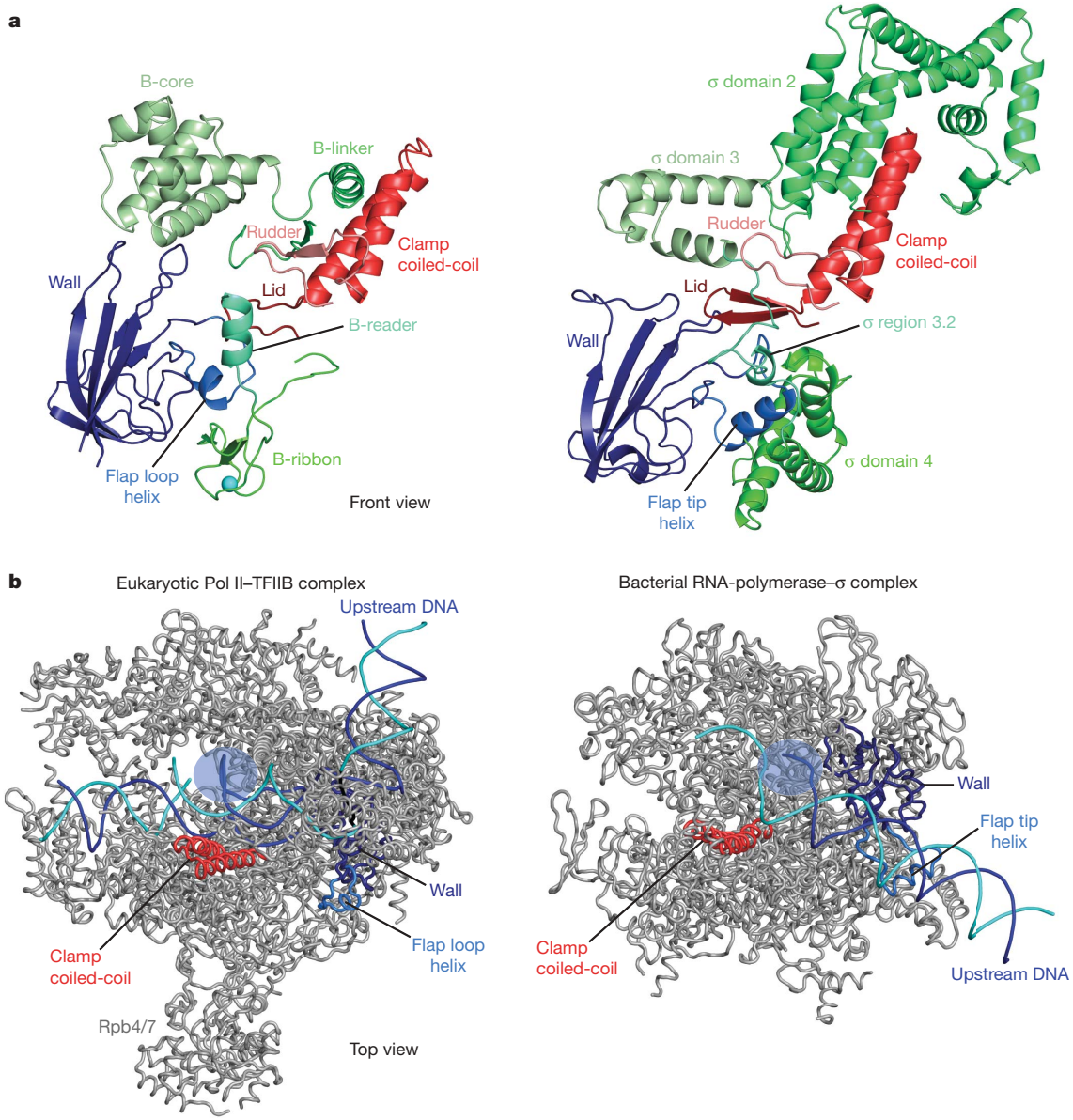


Figure 6 | Eukaryotic and bacterial transcription initiation complexes. **a**, Comparison of the polymerase-bound structures of B (left) and the bacterial initiation factor σ70 (ref. 44) (right). The view is from the front. The B-linker, B-core, B-ribbon and B-reader correspond topologically to parts of domains σ_2, σ_3, σ_4 and the σ3.2 linker, respectively, and are coloured accordingly. **b**, Comparison of the open complex model (left) with the structure of the bacterial RNA-polymerase–σ–upstream-DNA complex[45] (right). The view is from the top. Upstream DNA extends in different directions, but the point of DNA melting (circle) is similar in both complexes with respect to the polymerase. The conserved clamp coiled-coil and wall of the polymerases are in red and dark blue, respectively.

in promoter opening. The rudder is required for promoter opening in bacteria[49] and for transcription in archaea[18]. A region of the largest bacterial polymerase subunit that contains the clamp coiled-coil and σ factor is alone able to melt a promoter[50]. Consistent with topologically similar mechanisms of promoter opening, the points of DNA melting are similar in our PIC model and the bacterial RNA-polymerase–σ–DNA complex, although the courses of upstream DNA differ (Fig. 6b).

Conclusions

Our results rationalize genetic and biochemical data on transcription initiation collected over the last two decades. They also provide models of the closed and open complexes, reveal a new function of B in DNA opening, suggest a six-step mechanism for transcription initiation that includes the initiation–elongation transition, and unravel similarities and differences between eukaryotic and prokaryotic initiation machineries. The results also provide the framework for investigating the mechanisms underlying the regulation of transcription initiation, which governs cellular gene expression.

METHODS SUMMARY

Saccharomyces cerevisiae ten-subunit Pol II was prepared as described[4]. Recombinant Rpb4/7–B fusion protein was purified after expression in *Escherichia coli*. A Pol II–B–TBP–nucleic-acid complex was assembled and purified by size exclusion. Crystals were grown by vapour diffusion at 20 °C using 800 mM sodium ammonium tartrate, 100 mM HEPES, pH 7.5, 5 mM dithiothreitol (DTT) as precipitant, and contained only the Pol II–B complex. Diffraction data to 4.3 Å resolution were collected under cryo conditions at the European Synchrotron Radiation Facility (ESRF) beamline ID29. The structure was solved by molecular replacement. *Pfu* endogenous and recombinant RNA polymerases, and recombinant *Pfu* TBP, TFB and TFE were purified as described[23], with modifications. DNA templates were generated as reported[18] and run-off transcription was as described[23]. Permanganate footprints were as described[18]. Yeast nuclear extracts were prepared from wild-type strain and

strain SHY245 (ref. 10) as described (http://www.fhcrc.org/labs/hahn). *In vitro* transcription was performed essentially as reported[10].

Received 24 August; accepted 1 October 2009.
Published online 9 October 2009.

1. Roeder, R. G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21,** 327–335 (1996).
2. Cramer, P. *et al.* Structure of eukaryotic RNA polymerases. *Annu. Rev. Biophys.* **37,** 337–352 (2008).
3. Hahn, S. Structure and mechanism of the RNA polymerase II transcription machinery. *Nature Struct. Mol. Biol.* **11,** 394–403 (2004).
4. Armache, K.-J., Kettenberger, H. & Cramer, P. Architecture of the initiation-competent 12-subunit RNA polymerase II. *Proc. Natl Acad. Sci. USA* **100,** 6964–6968 (2003).
5. Bushnell, D. A. & Kornberg, R. D. Complete RNA polymerase II at 4.1-Å resolution: implications for the initiation of transcription. *Proc. Natl Acad. Sci. USA* **100,** 6969–6973 (2003).
6. Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Å resolution. *Science* **292,** 1876–1882 (2001).
7. Bushnell, D. A., Westover, K. D., Davis, R. E. & Kornberg, R. D. Structural basis of transcription: an RNA polymerase II–TFIIB cocrystal at 4.5 Angstroms. *Science* **303,** 983–988 (2004).
8. Chen, H. T. & Hahn, S. Binding of TFIIB to RNA polymerase II: mapping the binding site for the TFIIB zinc ribbon domain within the preinitiation complex. *Mol. Cell* **12,** 437–447 (2003).
9. Nikolov, D. B. *et al.* Crystal structure of a TFIIB–TBP–TATA-element ternary complex. *Nature* **377,** 119–128 (1995).
10. Ranish, J. A., Yudkovsky, N. & Hahn, S. Intermediates in formation and activity of the RNA polymerase II preinitiation complex: holoenzyme recruitment and a postrecruitment role for the TATA box and TFIIB. *Genes Dev.* **13,** 49–63 (1999).
11. Pardee, T. S., Bangur, C. S. & Ponticelli, A. S. The N-terminal region of yeast TFIIB contains two adjacent functional domains involved in stable RNA polymerase II binding and transcription start site selection. *J. Biol. Chem.* **273,** 17859–17864 (1998).
12. Cho, E. J. & Buratowski, S. Evidence that transcription factor IIB is required for a post-assembly step in transcription initiation. *J. Biol. Chem.* **274,** 25807–25813 (1999).
13. Chen, H. T. & Hahn, S. Mapping the location of TFIIB within the RNA polymerase II transcription preinitiation complex: a model for the structure of the PIC. *Cell* **119,** 169–180 (2004).
14. Edwards, A. M., Kane, C. M., Young, R. A. & Kornberg, R. D. Two dissociable subunits of yeast RNA polymerase II stimulate the initiation of transcription at a promoter *in vitro*. *J. Biol. Chem.* **266,** 71–75 (1991).
15. Armache, K.-J., Mitterweger, S., Meinhart, A. & Cramer, P. Structures of complete RNA polymerase II and its subcomplex Rpb4/7. *J. Biol. Chem.* **280,** 7131–7134 (2005).
16. Zhu, W. *et al.* The N-terminal domain of TFIIB from *Pyrococcus furiosus* forms a zinc ribbon. *Nature Struct. Biol.* **3,** 122–124 (1996).
17. Pinto, I., Wu, W.-H., Na, J. G. & Hampsey, M. Characterization of *sua7* mutations defines a domain of TFIIB involved in transcripiton start site selection in yeast. *J. Biol. Chem.* **269,** 30569–30573 (1994).
18. Naji, S., Bertero, M. G., Spitalny, P., Cramer, P. & Thomm, M. Structure-function analysis of the RNA polymerase cleft loops elucidates initial transcription, DNA unwinding and RNA displacement. *Nucleic Acids Res.* **36,** 676–687 (2008).
19. Kettenberger, H., Armache, K.-J. & Cramer, P. Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS. *Mol. Cell* **16,** 955–965 (2004).
20. Giardina, C. & Lis, J. T. DNA melting on yeast RNA polymerase II promoters. *Science* **261,** 759–762 (1993).
21. Renfrow, M. B. *et al.* Transcription factor B contacts promoter DNA near the transcription start site of the archaeal transcription initiation complex. *J. Biol. Chem.* **279,** 2825–2831 (2004).
22. Bartlett, M. S., Thomm, M. & Geiduschek, E. P. Topography of the euryarchaeal transcription initiation complex. *J. Biol. Chem.* **279,** 5894–5903 (2004).
23. Naji, S., Grunberg, S. & Thomm, M. The RPB7 orthologue E′ is required for transcriptional activity of a reconstituted archaeal core enzyme at low temperatures and stimulates open complex formation. *J. Biol. Chem.* **282,** 11047–11057 (2007).
24. Kusser, A. G. *et al.* Structure of an archaeal RNA polymerase. *J. Mol. Biol.* **376,** 303–307 (2008).
25. Hirata, A., Klein, B. J. & Murakami, K. S. The X-ray crystal structure of RNA polymerase from Archaea. *Nature* **451,** 851–854 (2008).
26. Kassavetis, G. A., Letts, G. A. & Geiduschek, E. P. The RNA polymerase III transcription initiation factor TFIIIB participates in two steps of promoter opening. *EMBO J.* **20,** 2823–2834 (2001).
27. Kuehner, J. N. & Brow, D. A. Quantitative analysis of *in vivo* initiator selection by yeast RNA polymerase II supports a scanning model. *J. Biol. Chem.* **281,** 14119–14128 (2006).
28. Bangur, C. S., Pardee, T. S. & Ponticelli, A. S. Mutational analysis of the D1/E1 core helices and the conserved N-terminal region of yeast transcription factor IIB (TFIIB): identification of an N-terminal mutant that stabilizes TATA-binding protein–TFIIB–DNA complexes. *Mol. Cell. Biol.* **17,** 6784–6793 (1997).
29. Faitar, S. L., Brodie, S. A. & Ponticelli, A. S. Promoter-specific shifts in transcription initiation conferred by yeast TFIIB mutations are determined by the sequence in the immediate vicinity of the start sites. *Mol. Cell. Biol.* **21,** 4427–4440 (2001).
30. Zhang, D. Y., Carson, D. J. & Ma, J. The role of TFIIB–RNA polymerase II interaction in start site selection in yeast cells. *Nucleic Acids Res.* **30,** 3078–3085 (2002).
31. Li, Y., Flanagan, P. M., Tschochner, H. & Kornberg, R. D. RNA polymerase II initiation factor interactions and transcription start site selection. *Science* **263,** 805–807 (1994).
32. Zhang, Z. & Dietrich, F. S. Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5′ SAGE. *Nucleic Acids Res.* **33,** 2838–2851 (2005).
33. Chen, B. S. & Hampsey, M. Functional interaction between TFIIB and the Rpb2 subunit of RNA polymerase II: implications for the mechanism of transcription initiation. *Mol. Cell. Biol.* **24,** 3983–3991 (2004).
34. Kwapisz, M. *et al.* Mutations of RNA polymerase II activate key genes of the nucleoside triphosphate biosynthetic pathways. *EMBO J.* **27,** 2411–2421 (2008).
35. Thiebaut, M. *et al.* Futile cycle of transcription initiation and termination modulates the response to nucleotide shortage in *S. cerevisiae*. *Mol. Cell* **31,** 671–682 (2008).
36. Kuehner, J. N. & Brow, D. A. Regulation of a eukaryotic gene by GTP-dependent start site selection and transcription attenuation. *Mol. Cell* **31,** 201–211 (2008).
37. Werner, F. & Weinzierl, R. O. Direct modulation of RNA polymerase core functions by basal transcription factors. *Mol. Cell. Biol.* **25,** 8344–8355 (2005).
38. Andrecka, J. *et al.* Nano positioning system reveals the course of upstream and nontemplate DNA within the RNA polymerase II elongation complex. *Nucleic Acids Res.* doi:10.1093/nar/gkp601 (20 July 2009).
39. Pal, M., Ponticelli, A. S. & Luse, D. S. The role of the transcription bubble and TFIIB in promoter clearance by RNA polymerase II. *Mol. Cell* **19,** 101–110 (2005).
40. Kim, T. K., Ebright, R. H. & Reinberg, D. Mechanism of ATP-dependent promoter melting by transcription factor IIH. *Science* **288,** 1418–1421 (2000).
41. Chen, H.-T., Warfield, L. & Hahn, S. The positions of TFIIF and TFIIE in the RNA polymerase II transcription initiation complex. *Nature Struct. Mol. Biol.* **14,** 696–703 (2007).
42. Cramer, P. Common structural features of nucleic acid polymerases. *Bioessays* **24,** 724–729 (2002).
43. Murakami, K. S., Masuda, S. & Darst, S. A. Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 Å resolution. *Science* **296,** 1280–1284 (2002).
44. Vassylyev, D. G. *et al.* Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å resolution. *Nature* **417,** 712–719 (2002).
45. Murakami, K. S., Masuda, S., Campbell, E. A., Muzzin, O. & Darst, S. A. Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. *Science* **296,** 1285–1290 (2002).
46. Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292,** 1863–1876 (2001).
47. Geszvain, K., Gruber, T. M., Mooney, R. A., Gross, C. A. & Landick, R. A hydrophobic patch on the flap-tip helix of *E. coli* RNA polymerase mediates σ^{70} region 4 function. *J. Mol. Biol.* **343,** 569–587 (2004).
48. Severinov, K. *et al.* The sigma subunit conserved region 3 is part of ''5′-face'' of active center of *Escherichia coli* RNA polymerase. *J. Biol. Chem.* **269,** 20826–20828 (1994).
49. Kuznedelov, K., Korzheva, N., Mustaev, A. & Severinov, K. Structure-based analysis of RNA polymerase function: the largest subunit's rudder contributes critically to elongation complex stability and is not involved in the maintenance of RNA–DNA hybrid length. *EMBO J.* **21,** 1369–1378 (2002).
50. Young, B. A., Gruber, T. M. & Gross, C. A. Minimal machinery of RNA polymerase holoenzyme sufficient for promoter melting. *Science* **303,** 1382–1384 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements P.C. was supported by the Deutsche Forschungsgemeinschaft, the Sonderforschungsbereich SFB646, the Transregio 5, the Forschergruppe 'Regulation und Mechanismen der Ribosomen-Biogenese', the Nanosystems Initiative Munich (NIM), the Ernst-Jung-Stiftung, and the Fonds der chemischen Industrie. M.S. was supported by the Boehringer-Ingelheim-Fonds and Elitenetzwerk Bayern. Part of this work was performed at the ESRF at Grenoble, France, and at the Swiss Light Source (SLS) at the Paul Scherrer Institut, Villigen, Switzerland. M.T. and M.E.Z. were supported by the DFG Forschergruppe 'Regulation und Mechanismen der Ribosomenbiogenese'. We thank members of the Cramer laboratory, in particular E. Lehmann and K. Maier, and members of the Thomm laboratory, in particular P. Decartes, W. Forster and F. Hirchaud. We thank K. Römer and the Römer-Stiftung for support.

Author Contributions D.K. carried out structure determination and modelling. M.E.Z. carried out archaeal biochemical assays. K.-J.A. prepared and measured the crystals and did the initial data processing. M.E.Z. and M.S. carried out yeast analysis. K.L. provided technical assistance. M.T. supervised the archaeal biochemical work. P.C. designed and supervised the project and prepared the manuscript.

Author Information Atomic coordinates and structure factors of the complete Pol II–B complex crystal structure have been deposited with the Protein Data Bank under accession number 3K1F. The closed and open complex models can be downloaded from http://www.lmb.uni-muenchen.de/cramer. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.C. (cramer@lmb.uni-muenchen.de).

LETTERS

A limit on the variation of the speed of light arising from quantum gravity effects

A list of authors and their affiliations appears at the end of the paper

A cornerstone of Einstein's special relativity is Lorentz invariance—the postulate that all observers measure exactly the same speed of light in vacuum, independent of photon-energy. While special relativity assumes that there is no fundamental length-scale associated with such invariance, there is a fundamental scale (the Planck scale, $l_{\rm Planck} \approx 1.62 \times 10^{-33}$ cm or $E_{\rm Planck} = M_{\rm Planck}c^2 \approx 1.22 \times 10^{19}$ GeV), at which quantum effects are expected to strongly affect the nature of space–time. There is great interest in the (not yet validated) idea that Lorentz invariance might break near the Planck scale. A key test of such violation of Lorentz invariance is a possible variation of photon speed with energy[1–7]. Even a tiny variation in photon speed, when accumulated over cosmological light-travel times, may be revealed by observing sharp features in γ-ray burst (GRB) light-curves[2]. Here we report the detection of emission up to ~31 GeV from the distant and short GRB 090510. We find no evidence for the violation of Lorentz invariance, and place a lower limit of $1.2E_{\rm Planck}$ on the scale of a linear energy dependence (or an inverse wavelength dependence), subject to reasonable assumptions about the emission (equivalently we have an upper limit of $l_{\rm Planck}/1.2$ on the length scale of the effect). Our results disfavour quantum-gravity theories[3,6,7] in which the quantum nature of space–time on a very small scale linearly alters the speed of light.

On 10 May 2009, at T_0 = 00:22:59.97 UT, both the Gamma-ray Burst Monitor (GBM)[8] and the Large Area Telescope (LAT)[9] onboard the Fermi Gamma-ray Space Telescope triggered on the very bright short GRB 090510 (hereafter all times are measured relative to T_0). Ground-based optical spectroscopy data, taken 3.5 days later[10], exhibited prominent emission lines at a common redshift of $z = 0.903 \pm 0.003$, corresponding to a luminosity distance of $d_{\rm L} = 1.8 \times 10^{28}$ cm (for a standard cosmology of $[\Omega_\Lambda, \Omega_{\rm M}, h] = [0.73, 0.27, 0.71]$). The GBM light curve (Fig. 1b, c; 8 keV–40 MeV) consists of seven main pulses. After the first dim short spike near trigger-time, the flux returns to background level; the main GBM emission starts at 0.53 s and lasts <0.5 s. The main LAT emission above 100 MeV starts at ~0.63 s and lasts <1 s with a decaying tail that extends to ~200 s.

A single 31-GeV photon was detected at 0.829 s, which coincides in time with the last of the seven GBM pulses (Fig. 1b, c, f). The nature of this Fermi/LAT event as a photon (rather than a background cosmic ray) was confirmed with very thorough analysis (see Supplementary Information section 1). We find the directional and temporal coincidence of this photon with GRB 090510 to be very significant, at $>5\sigma$ confidence, and find the 1σ confidence interval for its energy to be 27.97–36.32 GeV.

The known distance[10] ($z = 0.903 \pm 0.003$) of GRB 090510 and the detection of >1 GeV photons less than a second from its onset allow us to constrain the possible variation of the speed of light with photon-energy (known as photon dispersion: one form of the Lorentz Invariance Violation, LIV). Some quantum-gravity theories[2,4,5] are consistent with the photon-propagation speed $v_{\rm ph}$ varying with photon-energy $E_{\rm ph}$, and becoming considerably different from the ordinary (or low-energy limit of) speed of light, $c \equiv v_{\rm ph}(E_{\rm ph} \to 0)$, near the Planck scale (when $E_{\rm ph}$ becomes comparable to $E_{\rm Planck} = M_{\rm Planck}c^2$). For $E_{\rm ph} \ll E_{\rm Planck}$, the leading term in a Taylor series expansion of the classical dispersion relation is $|v_{\rm ph}/c - 1| \approx (E_{\rm ph}/M_{{\rm QG},n}c^2)^n$, where $M_{{\rm QG},n}$ is the quantum gravity mass for order n and $n = 1$ or 2 is usually assumed. The linear case ($n = 1$) gives a difference $\Delta t = \pm(\Delta E/M_{{\rm QG},1}c^2)D/c$ in the arrival time of photons emitted together at a distance D from us, and differing by $\Delta E = E_{\rm high} - E_{\rm low}$. At cosmological distances this simple expression is somewhat modified (see Supplementary Information section 4).

Because of their short duration (typically with short substructure consisting of pulses or narrow spikes) and cosmological distances, GRBs are well-suited for constraining LIV[2,11,12]. Individual spikes in long[13] (of duration >2 s) GRB light-curves (10–1,000 keV) usually show[14] intrinsic lags: the peak of a spike occurs earlier at higher photon-energies. However, there are either no lags or very short lags of either sign for short GRBs[15]. Thus far, intrinsic lags have been seen only on timescales of up to the width of individual spikes in a light curve, which for GRB 090510 are $\sim 10^{-2}$ s. Intrinsic lags have not yet been measured at high energies; if they are also present there, it is reasonable to assume that their behaviour is similar to that at low-energies (at least approximately).

When allowing for LIV-induced time-delays, the measured arrival time, $t_{\rm h}$, of the high-energy photons might not directly reflect their emission time, $t_{\rm em}$ (which would have been their arrival time if $v_{\rm ph} = c$). Therefore, we make reasonable and conservative assumptions on $t_{\rm em}$, constraining it using the observed lower-energy emission (for which LIV-induced time-delays are relatively negligible).

Using the DisCan method[12], we have searched for time delays within the LAT data (actual energy range of the photons used: 35 MeV–31 GeV) in the burst interval with the most intense emission (0.50–1.45 s). This approach extracts dispersion information from all detected LAT photons, and does not involve binning in either time or energy. It moves each photon to the time at which it would have been detected in the absence of any LIV-induced lag, given a trial value of the energy-lag coefficient. The value of this coefficient that maximizes a measure of the sharpness of the resulting light curve is an estimate of the apparent dispersion. Bootstrap error analysis[16] shows that this is not a detection, just an upper limit. For reasons similar to those advanced above (improbability of inherent lags or fortuitous cancellation of quantum gravity and intrinsic dispersion) we take this as an upper limit on LIV-induced dispersion. A similar method was described in ref. 17. We obtain a robust upper limit of $|\Delta t/\Delta E| < 30$ ms GeV^{-1} (at the 99% confidence level) on possible linear energy dispersion of either sign, or $\xi_1 \equiv M_{{\rm QG},1}/M_{\rm Planck} > 1.22$ (limit a in Table 1).

Using a different approach, we derive additional limits. To constrain a positive time delay ($v_{\rm ph} < c$, implying $t_{\rm h} > t_{\rm em}$) we do not attempt to associate the relevant high-energy photon with a particular spike in the low-energy light-curve. Instead, we simply assume that it was emitted sometime during the relevant lower-energy emission episode, that is, after its starting time $t_{\rm start}$ ($t_{\rm em} > t_{\rm start}$; see Fig. 1).

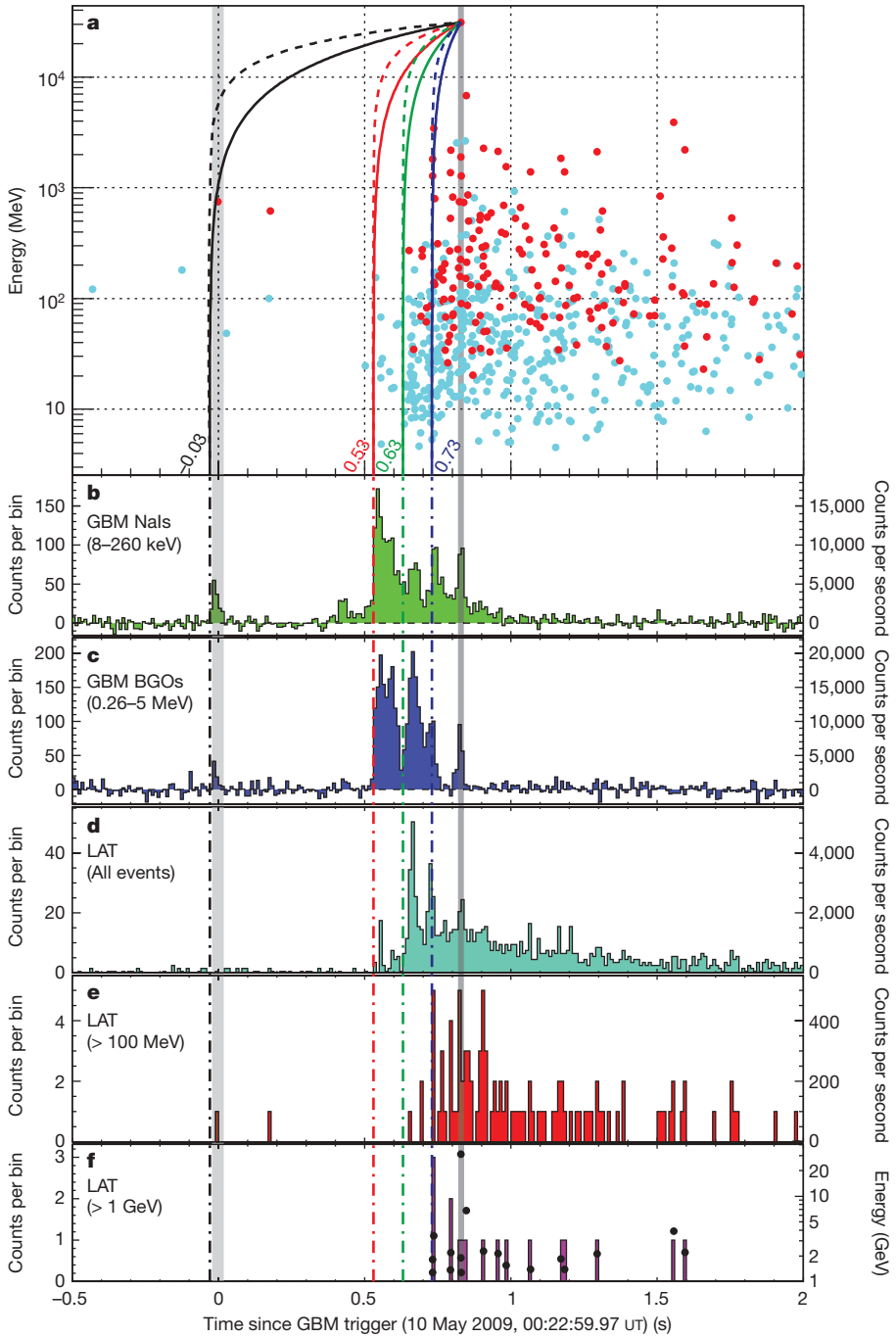


Figure 1 | Light curves of GRB 090510 at different energies. **a**, Energy versus arrival time with respect to the GBM trigger time for the 160 LAT photons that passed the transient off-line event selection (red) and the 161 photons that passed the onboard γ-ray filter (blue), and are consistent with the direction of GRB 090510. The solid and dashed curves are normalized to pass through the highest energy (31 GeV) photon and represent the relation between a photon's energy and arrival time for linear ($n = 1$) and quadratic ($n = 2$) LIV, respectively, assuming it is emitted at $t_{start} = -30$ ms (black; first small GBM pulse onset), 530 ms (red; main <1 MeV emission onset), 630 ms (green; >100 MeV emission onset), 730 ms (blue; >1 GeV emission onset). Photons emitted at t_{start} would be located along such a line owing to (a positive) LIV-induced time delay. **b**–**f**, GBM and LAT light curves, from lowest to highest energies. **f** also overlays energy versus arrival time for each photon, with the energy scale displayed on the right side. The dashed-dotted vertical lines show our four different possible choices for t_{start}. The grey shaded regions indicate the arrival time of the 31-GeV photon ±10 ms (on the right) and of a 750-MeV photon (during the first GBM pulse) ±20 ms (on the left), which can both constrain time delays of either sign. **b** and **c** show background-subtracted light curves for GBM NaI in the 8–260-keV band and a GBM BGO in the 0.260–5-MeV band, respectively. **d**, LAT events passing the onboard γ-ray filter. **e**, LAT transient class events with $E > 100$ MeV. **f**, LAT transient class events with $E > 1$ GeV. In all light curves, the time-bin width is 10 ms. In **b**–**e** the per-second count rate is displayed on the right for convenience.

This implies $\Delta t < t_h - t_{start}$ and thus sets a lower limit on M_{QG}. We have conservatively used the 31-GeV photon, even if another photon gave a stricter limit, because it is less sensitive to the exact choice of t_{start} or to intrinsic lags. In the following, we describe several possible different assumptions along with the astrophysical reasoning behind them and the corresponding lower limits on M_{QG}, starting from the most conservative assumption, and ending with the least conservative assumption (which is still very likely, and with good astrophysical motivation).

No high-energy photon has ever been detected before the onset of the low-energy emission in a GRB. Therefore, it is highly unlikely that the 31-GeV photon was emitted before the observed onset of

Table 1 | Limits on Lorentz invariance violation

	Limit on $\lvert\Delta t/\Delta E\rvert$ or $\lvert\Delta t\rvert$	Limit on $M_{QG,1}/M_{Planck}$	Valid for s_n
Limit a:	$\lvert\Delta t/\Delta E\rvert < 30\,\mathrm{ms\,GeV^{-1}}$	>1.22	±1
Limit b:	$\lvert\Delta t\rvert < 859\,\mathrm{ms}$	>1.19	1

Details for the derivations of these limits are given in the main text and Supplementary Information Section 4. Limit a is obtained by testing for an energy-dispersion in the high-energy (LAT all-event) data (that might smear the sharp observed spikes in the light curve); we find an upper limit for a linear dispersion of photons above 30 MeV of $\lvert\Delta t/\Delta E\rvert < 30\,\mathrm{ms\,GeV^{-1}}$ (at 99% confidence; see Supplementary Information section 3). Limit b relies on the 31-GeV photon, and conservatively uses the 1σ lower limit on its energy (28.0 GeV) and the 1σ lower limit on the redshift ($z = 0.900$). Limit b assumes that the 31-GeV photon was not emitted before the onset of any emission detected by Fermi, so that t_{start} is set to the onset of the first small isolated GRB spike, 30 ms before the GBM trigger time. $s_n = 1$ indicates a positive ($v_{ph} < c$) time-delay and $s_n = -1$ indicates a negative ($v_{ph} > c$) time-delay.

GRB 090510. This implies $\xi_1 > 1.19$ (limit b in Table 1), which we consider our most conservative limit with this method. While the underlying assumption on t_{em} is very reasonable, it is still an assumption; if for some reason t_{em} were to be before t_{start}, this limit would be weakened by a factor of $(t_h - t_{em})/(t_h - t_{start})$.

We stress here that our most conservative limits, a and b in Table 1, rely on very different and largely independent analysis, yet still give a very similar limit, of $\xi_1 > 1.2$. This lends considerable support to this result, and makes it more robust and secure than for each of the methods separately.

Our data can be used to set additional limits, which, although not as secure as the one mentioned above, are still very useful. Using the same approach as for limit b, we note that for a reasonable emission spectrum the 31-GeV photon would be accompanied by a large number of detectable (by either GBM or LAT) lower-energy photons, which suffer a much smaller LIV-induced time-delay, and thus 'mark' its emission time, $t_{em,31}$. If $t_{em,31}$ were during the first isolated GBM spike, then lower-energy photons emitted together with it should have been clustered near the black line in Fig. 1a owing to LIV-induced energy dispersion. Because this is not observed, it is much more likely that $t_{em,31}$ is associated with a later lower-energy emission episode. Setting t_{start} to the onset of the main GBM emission (530 ms) results in $\xi_1 > 3.42$.

Similarly, the expected large number of detectable >0.1-GeV photons emitted together with the 31-GeV photon makes it reasonable to set t_{start} to the onset of the main >0.1-GeV emission (630 ms; see Supplementary Information section 2), resulting in $\xi_1 > 5.12$. Correspondingly, the expected fair number of detectable >1-GeV photons emitted together with the 31-GeV photon makes it reasonable to set t_{start} to the onset of the main >1-GeV emission (730 ms), resulting in $\xi_1 > 10.0$. The $\xi_1 > 10.0$ value might be somewhat affected by the relatively small-number statistics for >1-GeV photons, or by intrinsic spectral lags (such effects are expected to be much smaller for the limit based on >0.1 GeV photons).

Finally, one can also set limits on LIV-induced time-delays of either sign based on the temporal association of the 31-GeV photon with the 7th GBM spike, and by associating the 0.75-GeV photon with the first GBM spike, because these photons arrive near the peak of a very narrow GBM spike (see Fig. 1), which is probably not due to chance coincidences. These associations would imply $\xi_1 > 102$ and 1.33, respectively. It is important to keep in mind, however, that while these associations are most likely, they are not very secure.

Our most secure and conservative new limit, $\xi_1 > 1.2$, is much stronger than the previous best limit of this kind ($\xi_1 > 0.1$ from GRB080916C; ref. 18) and fundamentally more meaningful. Given that in most quantum gravity scenarios $M_{QG,n} \leq M_{Planck}$, even our most conservative limits greatly reduce the parameter space for $n = 1$ models[19,20]. Our other limits, and especially our least conservative limit of $\xi_1 > 102$, make such theories highly implausible (models with $n > 1$ are not significantly constrained by our results). Thus, it is unlikely that other predictions of such $n = 1$ models would be observed. These include, for example, a reduction in the absorption of ≥10 TeV γ-rays by $\gamma\gamma \rightarrow e^+e^-$ interactions with extragalactic infrared photons[21,22], and fuzziness of radio or optical images of distant extragalactic sources[23–25]. Our stringent photon dispersion limit strongly disfavours models of Planck scale physics in which the quantum nature of space–time causes a linear variation of the speed of light with photon energy.

Received 12 August; accepted 12 October 2009.
Published online 28 October 2009.

1. Wheeler, J. A. & Ford, K. W. *Geons, Black Holes, and Quantum Foam: A Life in Physics* (W. W. Norton and Company, 1998).
2. Amelino-Camelia, G., Ellis, J., Mavromatos, N. E., Nanopoulos, D. V. & Sarkar, S. Tests of quantum gravity from observations of gamma-ray bursts. *Nature* **393,** 763–765 (1998).
3. Mattingly, D. Modern tests of Lorentz invariance. *Living Rev. Relativity* **8,** 5–84 (2005).
4. Kostelecky, V. A. & Mewes, M. Astrophysical tests of Lorentz and CPT violation with photons. *Astrophys. J.* **689,** L1–L4 (2008).
5. Amelino-Camelia, G. & Smolin, L. Prospects for constraining quantum gravity dispersion with near term observations. Preprint at ⟨http://arxiv.org/abs/0906.3731⟩ (2009).
6. Jacobson, T., Liberati, S. & Mattingly, D. Lorentz violation at high energy: concepts, phenomena and astrophysical constraints. *Ann. Phys.* **321,** 150–196 (2006).
7. Amelino-Camelia, G. Quantum gravity phenomenology. Preprint at ⟨http://arxiv.org/abs/0806.0339⟩ (2008).
8. Guiriec, S. *et al.* GRB 090510: Fermi GBM detection. *GCN Circ.* **9336** (2009).
9. Ohno, M. *et al.* Fermi LAT detection of GRB 090510. *GCN Circ.* **9334** (2009).
10. Rau, A. *et al.* GRB090510: VLT/FORS2 spectroscopic redshift. *GCN Circ.* **9353** (2009).
11. Rodríguez Martínez, M., Piran, T. & Oren, Y. GRB 051221A and Tests of Lorentz Symmetry. *J. Cosmol. Astroparticle Phys.* **05,** 017–023 (2006).
12. Scargle, J. D., Norris, J. P. & Bonnell, J. T. An algorithm for detecting quantum gravity photon dispersion in gamma-ray bursts: DisCan. *Astrophys. J.* **673,** 972–980 (2008).
13. Kouveliotou, C. *et al.* Identification of two classes of gamma-ray bursts. *Astrophys. J.* **413,** L101–L104 (1993).
14. Norris, J. P. *et al.* Spectral evolution in gamma-ray bursts. *Adv. Space Res.* **6,** 19–22 (1986).
15. Norris, J. P. & Bonnell, J. T. Short GRBs with extended emission. *Astrophys. J.* **643,** 266–275 (2006).
16. Efron, B. & Tibshirani, R. *An Introduction to the Bootstrap* (Chapman and Hall, 1993).
17. Albert, J. *et al.* Probing quantum gravity using photons from a flare of the active galactic nucleus Markarian 501 observed by the MAGIC telescope. *Phys. Lett. B* **668,** 253–257 (2008).
18. Abdo, A. A. *et al.* Fermi observations of high-energy gamma-ray emission from GRB 080916C. *Science* **323,** 1688–1693 (2009).
19. Ellis, J., Mavromatos, N. E. & Nanopoulos, D. V. Derivation of a vacuum refractive index in a stringy space time foam model. *Phys. Lett. B* **665,** 412–417 (2008).
20. Zloshchastiev, K. G. Logarithmic nonlinearity in theories of quantum gravity: origin of time and observational consequences. Preprint at ⟨http://arxiv.org/abs/0906.4282⟩ (2009).
21. Kifune, T. Invariance violation extends the cosmic-ray horizon? *Astrophys. J.* **518,** L21–L24 (1999).
22. Jacob, U. & Piran, T. Inspecting absorption in the spectra of extra-galactic gamma-ray sources for insight into Lorentz invariance violation. *Phys. Rev. D* **78,** 124010 (2008).
23. Christiansen, W. A. Jack Ng, Y. & van Dam, H. Probing spacetime foam with extragalactic sources. *Phys. Rev. Lett.* **96,** 051301 (2006).
24. Jack Ng, Y. Spacetime foam and dark energy. *AIP Conf. Proc.* **1115,** 74–79 (2009).
25. Jack Ng, Y. Spacetime foam: from entropy and holography to infinite statistics and nonlocality. Preprint at ⟨http://arxiv.org/abs/0801.2962⟩ (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The Fermi LAT Collaboration acknowledges support from a number of agencies and institutes for both the development and the operation of the LAT as well as scientific data analysis. These include NASA and DOE in the United States, CEA/Irfu and IN2P3/CNRS in France, ASI and INFN in Italy, MEXT, KEK, and JAXA in Japan, and the K. A. Wallenberg Foundation, the Swedish Research Council and the National Space Board in Sweden. Additional support from INAF in Italy for science analysis during the operations phase is also acknowledged. J. Granot gratefully acknowledges a Royal Society Wolfson Research Merit Award. The Fermi GBM Collaboration acknowledges the support of NASA in the United States and DRL in Germany. J. Conrad is a Royal Swedish Academy of Sciences Research Fellow, funded by a grant from the K. A. Wallenberg Foundation. E.T. is a NASA Postdoctoral Program Fellow and a Canon Foundation in Europe Fellow. A. J.v.d.H. is a NASA Postdoctoral Program Fellow. We thank J. Ellis for comments.

Author Contributions All authors contributed extensively to the work presented in this paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J. Granot (j.granot@herts.ac.uk), S. Guiriec (sylvain.guiriec@nasa.gov), M. Ohno (ohno@astro.isas.jaxa.jp) and V. Pelassa (pelassa@lpta.in2p3.fr).

A. A. Abdo[1,2], M. Ackermann[3], M. Ajello[3], K. Asano[4,5], W. B. Atwood[6], M. Axelsson[8,9], L. Baldini[12], J. Ballet[13], G. Barbiellini[14,15], M. G. Baring[16], D. Bastieri[17,18], K. Bechtol[3], R. Bellazzini[12], B. Berenji[3], P. N. Bhat[19], E. Bissaldi[20], E. D. Bloom[3], E. Bonamente[21,22], J. Bonnell[24,25], A. W. Borgland[3], A. Bouvier[3], J. Bregeon[12], A. Brez[12], M. S. Briggs[19], M. Brigida[26,27], P. Bruel[28], J. M. Burgess[19], T. H. Burnett[29], G. A. Caliandro[26,27], R. A. Cameron[3], P. A. Caraveo[30], J. M. Casandjian[13], C. Cecchi[21,22], Ö. Çelik[23,24,31], V. Chaplin[19], E. Charles[3], C. C. Cheung[1,2,24], J. Chiang[3], S. Ciprini[21,22], R. Claus[3], J. Cohen-Tanugi[32], L. R. Cominsky[33], V. Connaughton[19], J. Conrad[9,10], S. Cutini[34], C. D. Dermer[1], A. de Angelis[35], F. de Palma[26,27], S. W. Digel[3], B. L. Dingus[36], E. do Couto e Silva[3], P. S. Drell[3], R. Dubois[3], D. Dumora[37], C. Farnier[32], C. Favuzzi[26,27], S. J. Fegan[28], J. Finke[1,2], G. Fishman[38], W. B. Focke[3], L. Foschini[39], Y. Fukazawa[40], S. Funk[3], P. Fusco[26,27], F. Gargano[27], D. Gasparrini[34], N. Gehrels[24,25], S. Germani[21,22], L. Gibby[41], B. Giebels[28], N. Giglietto[26,27], F. Giordano[26,27], T. Glanzman[3], G. Godfrey[3], J. Granot[42], J. Greiner[20], I. A. Grenier[13], M.-H. Grondin[37], J. E. Grove[1], D. Grupe[43], L. Guillemot[44], S. Guiriec[19], Y. Hanabata[40], A. K. Harding[24], M. Hayashida[3], E. Hays[24], E. A. Hoversten[43], R. E. Hughes[45], G. Jóhannesson[3], A. S. Johnson[3], R. P. Johnson[6], W. N. Johnson[1], T. Kamae[3], H. Katagiri[40], J. Kataoka[4,46], N. Kawai[4,47], M. Kerr[29], R. M. Kippen[36], J. Knödlseder[48], D. Kocevski[3], C. Kouveliotou[38], F. Kuehn[45], M. Kuss[12], J. Lande[3], L. Latronico[12], M. Lemoine-Goumard[37], F. Longo[14,15], F. Loparco[26,27], B. Lott[37], M. N. Lovellette[1], P. Lubrano[21,22], G. M. Madejski[3], A. Makeev[1,49], M. N. Mazziotta[27], S. McBreen[20,50], J. E. McEnery[24], S. McGlynn[9,11], P. Mészáros[43], C. Meurer[9,10], P. F. Michelson[3], W. Mitthumsiri[3], T. Mizuno[40], A. A. Moiseev[25,23], C. Monte[26,27], M. E. Monzani[3], E. Moretti[14,15], A. Morselli[51], I. V. Moskalenko[3], S. Murgia[3], T. Nakamori[4], P. L. Nolan[3], J. P. Norris[53], E. Nuss[32], M. Ohno[54], T. Ohsugi[40], N. Omodei[12], E. Orlando[20], J. F. Ormes[53], M. Ozaki[54], W. S. Paciesas[19], D. Paneque[3], J. H. Panetta[3], D. Parent[37], V. Pelassa[32], M. Pepe[21,22], M. Pesce-Rollins[12], V. Petrosian[3], F. Piron[32], T. A. Porter[6], R. Preece[19], S. Rainò[26,27], E. Ramirez-Ruiz[7], R. Rando[17,18], M. Razzano[12], S. Razzaque[1,2], A. Reimer[3,55], O. Reimer[3,55], T. Reposeur[37], S. Ritz[6], L. S. Rochester[3], A. Y. Rodriguez[56], M. Roth[29], F. Ryde[9,11], H. F.-W. Sadrozinski[6], D. Sanchez[28], A. Sander[45], P. M. Saz Parkinson[6], J. D. Scargle[57], T. L. Schalk[6], C. Sgrò[12], E. J. Siskind[58], D. A. Smith[37], P. D. Smith[45], G. Spandre[12], P. Spinelli[26,27], M. Stamatikos[24,45], F. W. Stecker[24], M. S. Strickman[1], D. J. Suson[59], H. Tajima[3], H. Takahashi[40], T. Takahashi[54], T. Tanaka[3], J. B. Thayer[3], J. G. Thayer[3], D. J. Thompson[24], L. Tibaldo[13,17,18], K. Toma[43], D. F. Torres[56,60], G. Tosti[21,22], E. Troja[4,24], Y. Uchiyama[3,54], T. Uehara[40], T. L. Usher[3], A. J. van der Horst[38], V. Vasileiou[24,23,31], N. Vilchez[48], V. Vitale[51,52], A. von Kienlin[20], A. P. Waite[3], P. Wang[3], C. Wilson-Hodge[38], B. L. Winer[45], K. S. Wood[1], X. F. Wu[43,61,62], R. Yamazaki[40], T. Ylinen[9,11,63] & M. Ziegler[6]

[1]Space Science Division, Naval Research Laboratory, Washington, District of Columbia 20375, USA. [2]National Research Council Research Associate, National Academy of Sciences, Washington, District of Columbia 20001, USA. [3]W. W. Hansen Experimental Physics Laboratory, Kavli Institute for Particle Astrophysics and Cosmology, Department of Physics and SLAC National Accelerator Laboratory, Stanford University, Stanford, California 94305, USA. [4]Department of Physics, [5]Interactive Research Center of Science, Tokyo Institute of Technology, Meguro City, Tokyo 152-8551, Japan. [6]Santa Cruz Institute for Particle Physics, Department of Physics and Department of Astronomy and Astrophysics, University of California at Santa Cruz, [7]UCO/Lick Observatories, Santa Cruz, California 95064, USA. [8]Department of Astronomy, Stockholm University, [9]The Oskar Klein Centre for Cosmoparticle Physics, [10]Department of Physics, [11]Department of Physics, Royal Institute of Technology (KTH), AlbaNova, SE-106 91 Stockholm, Sweden. [12]Istituto Nazionale di Fisica Nucleare, Sezione di Pisa, I-56127 Pisa, Italy. [13]Laboratoire AIM, CEA-IRFU/CNRS/Université Paris Diderot, Service d'Astrophysique, CEA Saclay, 91191 Gif-sur-Yvette, France. [14]Istituto Nazionale di Fisica Nucleare, Sezione di Trieste, [15]Dipartimento di Fisica, Università di Trieste, I-34127 Trieste, Italy. [16]Rice University, Department of Physics and Astronomy, MS-108, P. O. Box 1892, Houston, Texas 77251, USA. [17]Istituto Nazionale di Fisica Nucleare, Sezione di Padova, [18]Dipartimento di Fisica "G. Galilei", Università di Padova, I-35131 Padova, Italy. [19]CSPAR, University of Alabama in Huntsville, Huntsville, Alabama 35899, USA. [20]Max-Planck Institut für extraterrestrische Physik, 85748 Garching, Germany. [21]Istituto Nazionale di Fisica Nucleare, Sezione di Perugia, [22]Dipartimento di Fisica, Università degli Studi di Perugia, I-06123 Perugia, Italy. [23]Center for Research and Exploration in Space Science and Technology (CRESST), [24]NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. [25]University of Maryland, College Park, Maryland 20742, USA. [26]Dipartimento di Fisica "M. Merlin" dell'Università e del Politecnico di Bari, [27]Istituto Nazionale di Fisica Nucleare, Sezione di Bari, I-70126 Bari, Italy. [28]Laboratoire Leprince-Ringuet, École polytechnique, CNRS/IN2P3, F-91128 Palaiseau, France. [29]Department of Physics, University of Washington, Seattle, Washington 98195-1560, USA. [30]INAF-Istituto di Astrofisica Spaziale e Fisica Cosmica, I-20133 Milano, Italy. [31]University of Maryland, Baltimore County, Baltimore, Maryland 21250, USA. [32]Laboratoire de Physique Théorique et Astroparticules, Université Montpellier 2, CNRS/IN2P3, 34095 Montpellier, Cedex 5, France. [33]Department of Physics and Astronomy, Sonoma State University, Rohnert Park, California 94928-3609, USA. [34]Agenzia Spaziale Italiana (ASI) Science Data Center, I-00044 Frascati (Roma), Italy. [35]Dipartimento di Fisica, Università di Udine and Istituto Nazionale di Fisica Nucleare, Sezione di Trieste, Gruppo Collegato di Udine, I-33100 Udine, Italy. [36]Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. [37]Université de Bordeaux and CNRS/IN2P3, Centre d'Études Nucléaires Bordeaux Gradignan, UMR 5797, Gradignan 33175, France. [38]Space Science Office, VP62, NASA/Marshall Space Flight Center, Huntsville, Alabama 35812, USA. [39]INAF Osservatorio Astronomico di Brera, I-23807 Merate, Italy. [40]Department of Physical Sciences, Hiroshima University, Higashi-Hiroshima, Hiroshima 739-8526, Japan. [41]Jacobs Technology, Huntsville, Alabama 35806, USA. [42]Centre for Astrophysics Research, University of Hertfordshire, College Lane, Hatfield AL10 9AB, UK. [43]Department of Astronomy and Astrophysics, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [44]Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, 53121 Bonn, Germany. [45]Department of Physics, Center for Cosmology and Astro-Particle Physics, The Ohio State University, Columbus, Ohio 43210, USA. [46]Waseda University, 1-104 Totsukamachi, Shinjuku-ku, Tokyo, 169-8050, Japan. [47]Cosmic Radiation Laboratory, Institute of Physical and Chemical Research (RIKEN), Wako, Saitama 351-0198, Japan. [48]Centre d'Étude Spatiale des Rayonnements, CNRS/UPS, BP 44346, F-30128 Toulouse cedex 4, France. [49]George Mason University, Fairfax, Virginia 22030, USA. [50]University College Dublin, Belfield, Dublin 4, Ireland. [51]Istituto Nazionale di Fisica Nucleare, Sezione di Roma "Tor Vergata", [52]Dipartimento di Fisica, Università di Roma "Tor Vergata", I-00133 Roma, Italy. [53]Department of Physics and Astronomy, University of Denver, Denver, Colorado 80208, USA. [54]Institute of Space and Astronautical Science, JAXA, 3-1-1 Yoshinodai, Sagamihara, Kanagawa 229-8510, Japan. [55]Institut für Astro- und Teilchenphysik and Institut für Theoretische Physik, Leopold-Franzens-Universität Innsbruck, A-6020 Innsbruck, Austria. [56]Institut de Ciencies de l'Espai (IEEC-CSIC), Campus UAB, 08193 Barcelona, Spain. [57]Space Sciences Division, NASA Ames Research Center, Moffett Field, California 94035-1000, USA. [58]NYCB Real-Time Computing Inc., Lattingtown, New York 11560-1025, USA. [59]Department of Chemistry and Physics, Purdue University Calumet, Hammond, Indiana 46323-2094, USA. [60]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08193 Barcelona, Spain. [61]Joint Center for Particle Nuclear Physics and Cosmology (J-CPNPC), [62]Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210008, China. [63]School of Pure and Applied Natural Sciences, University of Kalmar, SE-391 82 Kalmar, Sweden.

LETTERS

Coexistence of Fermi arcs and Fermi pockets in a high-T_c copper oxide superconductor

Jianqiao Meng[1], Guodong Liu[1], Wentao Zhang[1], Lin Zhao[1], Haiyun Liu[1], Xiaowen Jia[1], Daixiang Mu[1], Shanyu Liu[1], Xiaoli Dong[1], Jun Zhang[1], Wei Lu[1], Guiling Wang[2], Yong Zhou[2], Yong Zhu[2], Xiaoyang Wang[2], Zuyan Xu[2], Chuangtian Chen[2] & X. J. Zhou[1]

In the pseudogap state of the high-transition-temperature (high-T_c) copper oxide superconductors[1], angle-resolved photoemission (ARPES) measurements have seen Fermi arcs—that is, open-ended gapless sections in the large Fermi surface[2–8]—rather than a closed loop expected of an ordinary metal. This is all the more puzzling because Fermi pockets (small closed Fermi surface features) have been suggested by recent quantum oscillation measurements[9–14]. The Fermi arcs cannot be understood in terms of existing theories, although there is a solution in the form of conventional Fermi surface pockets associated with competing order, but with a back side that is for detailed reasons invisible to photoemission probes[15]. Here we report ARPES measurements of $Bi_2Sr_{2-x}La_xCuO_{6+\delta}$ (La-Bi2201) that reveal Fermi pockets. The charge carriers in the pockets are holes, and the pockets show an unusual dependence on doping: they exist in underdoped but not overdoped samples. A surprise is that these Fermi pockets appear to coexist with the Fermi arcs. This coexistence has not been expected theoretically.

The high-resolution Fermi surface mapping (Fig. 1a) of the underdoped La-Bi2201 sample (with T_c=18 K: UD18K) using a vacuum ultraviolet (VUV) laser reveals three Fermi surface sheets with low spectral weight (labelled LP, LS and LPS in Fig. 1a) in the momentum space covered, in addition to the prominent main Fermi surface (LM). One particular Fermi surface sheet LP crosses the main band LM, forming an enclosed loop—a Fermi pocket—near the nodal region. Quantitative Fermi surface data measured using both a VUV laser (Fig. 2a) and a helium discharge lamp (Fig. 2b) are summarized in Fig. 2c. All the possible umklapp bands[16,17], shadow bands[18,19] and umklapp bands of the shadow bands that may be present in the bismuth-based compounds are shown for comparison (see Supplementary Information and Supplementary Fig. 2 for more details). It is clear that the LP and LPS bands observed in VUV laser measurements (Fig. 2a) and the HP band observed in helium lamp measurements (Fig. 2b) are intrinsic; they cannot be attributed to any of the umklapp bands or shadow bands (Fig. 2c). The location of the three bands can be well connected by the same superlattice vector, indicating that the HP and LPS bands correspond to the first-order umklapp bands of the main Fermi pocket, LP. The shape and area of the Fermi pockets are also consistent in these two independent measurements, making a convincing case for the presence of the Fermi pocket. We note that in both the laser (Fig. 2a) and helium lamp

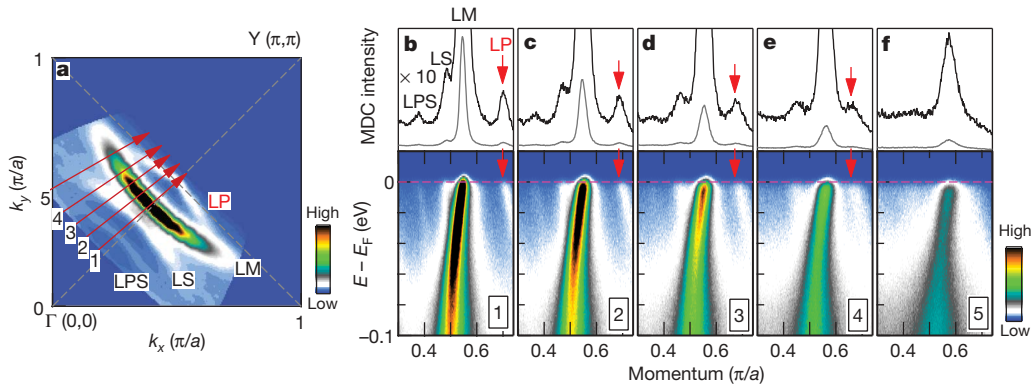


Figure 1 | Fermi surface and band structure of a La-Bi2201 sample. In the present paper, we use the phrase 'Fermi surface' loosely to denote the momentum space locus of high-intensity low-energy spectral weight. In copper oxide superconductors, the Fermi surface is usually measured in the superconducting state in spite of the gap opening because the sharpness of the features at low temperature facilitates the precise determination of the underlying Fermi surface as compared to the normal state. It has been shown that the 'underlying Fermi surface' determined from the minimum gap locus in the superconducting state is identical to that in the normal state[31]. **a**, Photoemission intensity at the Fermi energy (E_F) as a function of momenta k_x and k_y for the La-Bi2201 UD18K sample (underdoped, T_c = 18 K) measured at a temperature of 14 K. It is obtained by symmetrizing the original data with respect to the (0,0)–(π,π) line. Four Fermi surface sheets are resolved in the covered momentum space, marked as LM for the main sheet, LP for the Fermi pocket, and LS and LPS for the others. **b**–**f**, Band structure (bottom panels) and corresponding momentum distribution curves (MDCs) at the Fermi level (upper panels) along five typical momentum cuts (cuts 1 to 5) as labelled in Fig. 1a. To see the weak features more clearly, the original MDCs (thin grey lines) in the upper panel are expanded 10 times and plotted in the same figures (thick black lines). Note that the signal of the Fermi pocket LP is very weak; its intensity is over one order of magnitude weaker than that of the main band LM.

[1]National Laboratory for Superconductivity, Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China. [2]Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Beijing 100190, China.

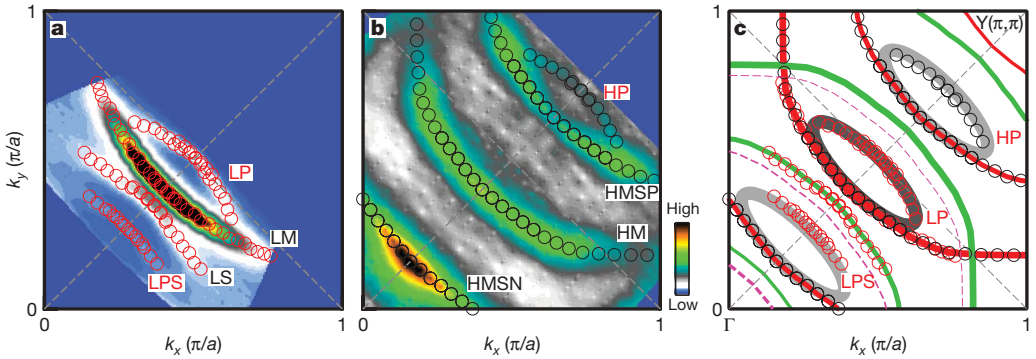


Figure 2 | Identification of the Fermi pocket in the photoemission data. **a**, Fermi surface measured in the La-Bi2201 UD18K sample using the VUV laser. The four observed Fermi surface sheets in the covered momentum space are quantitatively represented by red circles. **b**, Fermi surface measured in the La-Bi2201 UD18K sample by a helium discharge lamp at a photon energy of 21.218 eV. It is obtained by symmetrizing the original data with respect to the (0,0)–(π,π) line. Black circles represent quantitative positions of the observed Fermi surface sheets. **c**, Summary of the measured Fermi surface from VUV laser (open red circles, as in **a**) and helium discharge lamp (open black circles, as in **b**) observations. Red lines represent the main Fermi surface (central thickest line) and its corresponding umklapp bands (thinner lines on either side of the main Fermi surface). The thickest green line represents the shadow band of the main Fermi surface, and thinner green lines the umklapp bands of the shadow band. The pink dashed lines represent possible high-order umklapp bands from the main band in the third quadrant (see Supplementary Information and Supplementary Fig. 2 for more details). It is clear that the main Fermi surface measured from the VUV laser (LM) shows a good agreement with that from the helium discharge lamp (HM). The LS band observed using the VUV laser agrees well with the first-order umklapp band of the shadow band. The HMSP and HMSN bands observed in helium discharge lamp measurements correspond to the first-order umklapp bands of the main band. The three ellipses represent the position of the observed Fermi pockets. The matrix element effect (associated with the photoemission process) on the relative spectral intensity between the main band and the Fermi pocket is an interesting issue to be further explored.

(Fig. 2b and Supplementary Fig. 2) measurements, all the observed bands except for the 'Fermi pocket bands' can be explained by only one regular superstructure wavevector (0.24,0.24). The presence of additional superstructure, which would give rise to new bands, appears to be unlikely because there is no indication of such additional bands observed in our measurements.

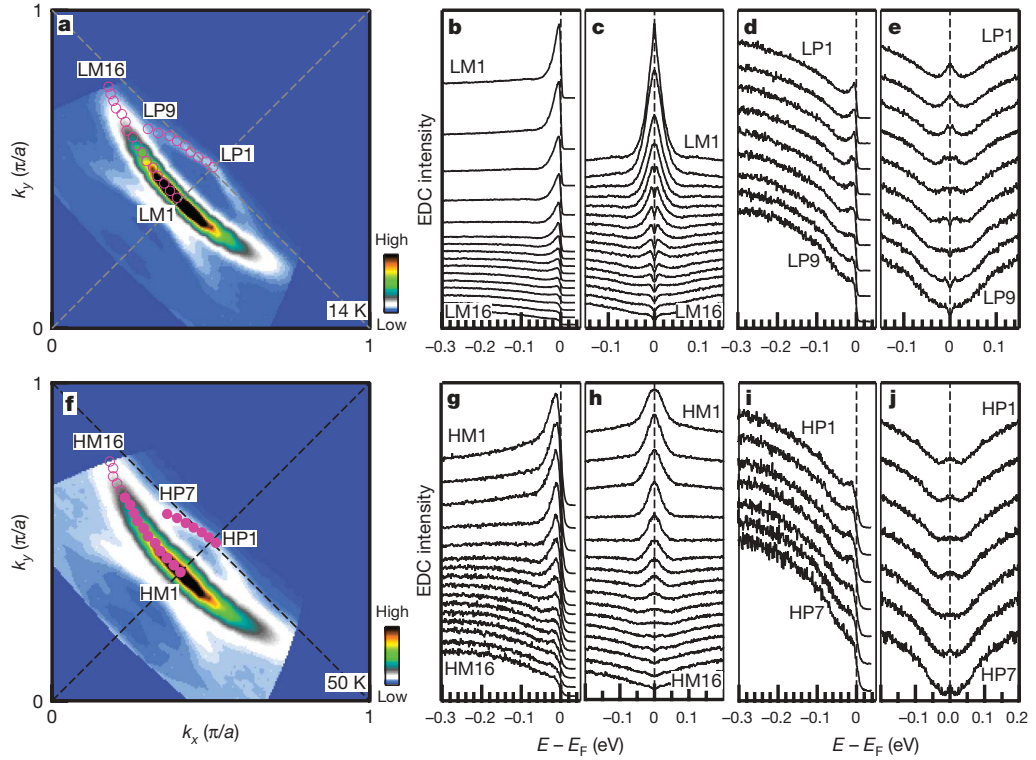


Figure 3 | Temperature dependence of the Fermi pocket. Top panels, Fermi surface and photoemission spectra for the La-Bi2201 UD18K sample below T_c; bottom panels, as in top panels but above T_c. **a**, Fermi surface mapping at 14 K. **b**, Photoemission spectra (energy distribution curves, EDCs) along the main Fermi surface. Sharp peaks are observed near the nodal region and get weaker when moving to the antinodal region. **c**, The corresponding symmetrized EDCs. The symmetrization procedure[32] provides an intuitive way to identify an energy gap opening which is characterized by the appearance of a dip near the Fermi level; zero gap corresponds to a peak at the Fermi level. The gap size is determined by the EDC peak position with respect to the Fermi level. It is clearly seen that the gap size increases from the nodal to antinodal regions. **d**, EDCs along the back side of the Fermi pocket. **e**, The corresponding symmetrized EDCs. The gap size increase is also clear going from the nodal to antinodal regions along the pocket. **f**, Fermi surface mapping at 50 K. **g**, EDCs along the main Fermi surface. **h**, The corresponding symmetrized EDCs. At 50 K, there is a gapless region formed near the nodal region with its location denoted in **f** by pink filled circles; the main Fermi surface beyond this region remains gapped. **i**, EDCs along the Fermi pocket. **j**, The corresponding symmetrized EDCs. At 50 K, the weak peak near the Fermi level indicates gap closing along the Fermi pocket.

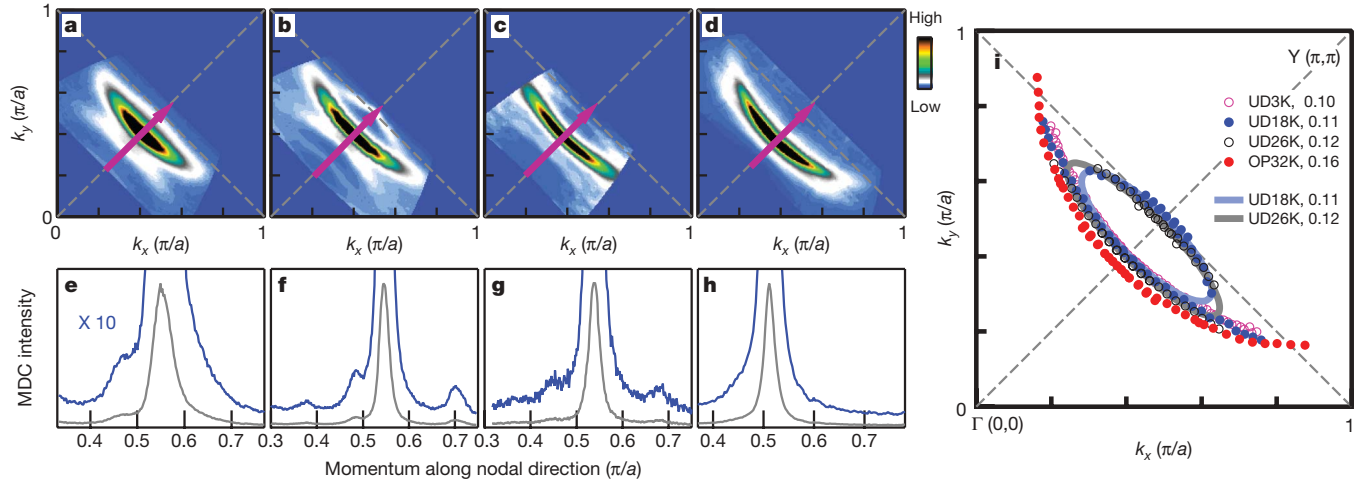


Figure 4 | Doping evolution of Fermi surface topology in La-Bi2201. **a**–**d**, Fermi surface mapping of UD3K (**a**), UD18K (**b**), UD26K (**c**) and OP32K[20] (**d**). These maps are obtained by symmetrizing the original data with respect to the (0,0)–(π,π) line. **e**–**h**, Corresponding MDCs at the Fermi energy along the (0,0)–(π,π) nodal direction. The location of the momentum cuts is labelled in **a**–**d** by purple lines with arrows. To show the weak features more clearly, the original MDCs (thin grey lines in bottom panels in **e**–**h**) are expanded by ten times and plotted in the same figures (blue thick lines). Note that we can not completely exclude the possibility of the presence of a Fermi pocket in the UD3K sample (**a**). It is noted that the MDC width of the sample shown in **e** is much greater than that of other dopings (**f**–**h**), leaving a possibility that the weak signal of a Fermi pocket may be buried in the broad feature. **i**, Quantitative Fermi surface of La-Bi2201 at various doping levels. The two 'Fermi pockets' for UD18K and UD26K samples are obtained by fitting both sides of the data points with ellipses.

The Fermi pocket is observed both in the normal state and the superconducting state, as shown in Fig. 3 for the La-Bi2201 UD18K sample. Moreover, its location, shape and area show little change with temperature (Fig. 3a and f). Below T_c, the opening of a superconducting gap is clearly observed on both the main band (Fig. 3c) and the back side of the Fermi pocket (Fig. 3e). Above T_c, a portion of the main Fermi surface near the nodal region becomes gapless, as indicated by the pink filled circles on the main Fermi surface (Fig. 3f). This is reminiscent of the Fermi arc formation in previous ARPES results, which show an increase in arc length with increasing temperature[3,5]. The back side of the Fermi pocket also becomes gapless above T_c (Fig. 3j), forming a gapless Fermi pocket loop with part of the main band. We note that the Fermi arc on the main band appears to extend farther than the Fermi pocket section (Fig. 3f), giving rise to an interesting coexistence of Fermi arc and Fermi pocket.

The Fermi pocket exhibits an unusual doping dependence, as seen in Fig. 4. It is observed not only in the UD18K sample (Fig. 4b), but also in the UD26K sample (Fig. 4c). But it is not seen in the UD3K sample (Fig. 4a) or the OP32K optimally doped sample (Fig. 4d)[20]. This peculiar doping dependence of the Fermi pocket—it can only be observed in a limited doping range in the underdoped region—lends further support to its intrinsic nature. The quantitative main Fermi surface and Fermi pocket at various dopings are summarized in Fig. 4i. The areas enclosed by the Fermi pockets in the UD18K (blue ellipse in Fig. 4i) and UD26K (grey ellipse in Fig. 4i) samples correspond to doping levels of 0.11 and 0.12 holes per copper site, respectively, which is in good agreement with the estimated hole concentration of each sample[21,22].

The Fermi pocket we have observed is hole-like, so it cannot correspond to the electron-like Fermi pocket suggested from quantum oscillation experiments[11,23]. Moreover, the observed Fermi pocket is not symmetrically located in the Brillouin zone: specifically, it is not centred around (π/2,π/2) (Fig. 4i). This is distinct from the observation reported in the Nd-LSCO system[24] which is symmetrical with respect to the (π,0)–(0,π) line, similar to the 'shadow band' commonly observed in Bi2201 and Bi2212[18,19]. This particular location makes it impossible that the Fermi pocket we have observed originates from the *d*-density-wave 'hidden order' that gives a hole-like Fermi pocket centred around the (π/2,π/2) point[23,25]. Among other possible origins of Fermi pocket formation[26–29], the phenomenological resonant valence bond picture[27,28] shows a fairly good agreement with our observations, in terms of the location, shape and area of the hole-like Fermi pocket and its doping dependence. In particular, the predicted Fermi pockets are pinned at the (π/2,π/2) point[27,28], which is consistent with our experiment (Fig. 4i). One obvious discrepancy is that the spectral weight on the back side of the Fermi pocket near (π/2,π/2) is expected to be zero from these theories[26–28], which is at odds with our measurements (Figs 3d and i). We note that the existence of incommensurate density waves could also potentially explain the observed pockets. One possible wavevector needed is $(1 \pm 0.092, 1 \pm 0.092)$, which is diagonal and can be examined by neutron or X-ray scattering measurements. There are indications of the charge-density-wave (CDW) formation reported in the Bi2201 system[30]; whether the Fermi surface reconstruction caused by such a CDW order or its related spin-density-wave order can account for our observation needs to be further explored.

One peculiar characteristic of the Fermi pocket is its coexistence with the large underlying Fermi surface (Figs 3 and 4). One may wonder whether this could originate from sample inhomogeneity: that is, the large Fermi surface is caused by a phase at high doping, whereas the Fermi pocket is from a phase at low doping. We believe this is unlikely, because there is no indication of two distinct superconducting phases being detected in the magnetization measurement (Supplementary Fig. 1). In addition, no Fermi pocket is identified for the underdoped UD3K sample (Fig. 4a). A more surprising observation is the coexistence of Fermi arcs and Fermi pockets in the normal state, because this has not been expected theoretically. We believe these new findings will provide key insights for understanding the anomalous pseudogap state in high-T_c copper oxide superconductors.

Received 14 May; accepted 10 September 2009.

1. Timusk, T. & Statt, B. The pseudogap in high-temperature superconductors: an experimental survey. *Rep. Prog. Phys.* **62**, 61–122 (1999).
2. Marshall, D. S. *et al.* Unconventional electronic structure evolution with hole doping in $Bi_2Sr_2CaCu_2O_{8+\delta}$: angle-resolved photoemission results. *Phys. Rev. Lett.* **76**, 4841–4844 (1996).
3. Norman, M. R. *et al.* Destruction of the Fermi surface in underdoped high-T_c superconductors. *Nature* **392**, 157–160 (1998).
4. Shen, K. M. *et al.* Nodal quasiparticles and antinodal charge ordering in $Ca_{2-x}Na_xCuO_2Cl_2$. *Science* **307**, 901–904 (2005).
5. Kanigel, A. *et al.* Evolution of the pseudogap from Fermi arcs to the nodal liquid. *Nature Phys.* **2**, 447–451 (2006).

6. Lee, W. S. *et al.* Abrupt onset of a second energy gap at the superconducting transition of underdoped Bi2212. *Nature* **450,** 81–84 (2007).
7. Hossain, M. A. *et al. In situ* doping control of the surface of high-temperature superconductors. *Nature Phys.* **4,** 527–531 (2008).
8. Yang, H. B. *et al.* Emergence of preformed Cooper pairs from the doped Mott insulating state in $Bi_2Sr_2CaCu_2O_{8+\delta}$. *Nature* **456,** 77–80 (2008).
9. Doiron-Leyraud, N. *et al.* Quantum oscillations and the Fermi surface in an underdoped high-T_c superconductor. *Nature* **447,** 565–568 (2007).
10. Bangura, A. *et al.* Small Fermi surface pockets in underdoped high temperature superconductors: observation of Shubnikov–de Haas oscillations in $YBa_2Cu_4O_8$. *Phys. Rev. Lett.* **100,** 047004 (2008).
11. LeBoeuf, D. *et al.* Electron pockets in the Fermi surface of hole-doped high-T_c superconductors. *Nature* **450,** 533–536 (2007).
12. Yelland, E. A. *et al.* Quantum oscillations in the underdoped cuprate $YBa_2Cu_4O_8$. *Phys. Rev. Lett.* **100,** 047003 (2008).
13. Jaudet, C. *et al.* de Haas–van Alphen oscillations in the underdoped high temperature superconductor $YBa_2Cu_3O_{6.5}$. *Phys. Rev. Lett.* **100,** 187005 (2008).
14. Sebastian, S. E. *et al.* A multi-component Fermi surface in the vortex state of an underdoped high-T_c superconductor. *Nature* **454,** 200–203 (2008).
15. Chakravarty, S., Nayak, C. & Tewari, S. Angle-resolved photoemission spectra in the cuprates from the d-density wave theory. *Phys. Rev. B* **68,** 100504 (2003).
16. Ding, H. *et al.* Momentum dependence of the superconducting gap in $Bi_2Sr_2CaCu_2O_8$. *Phys. Rev. Lett.* **74,** 2784–2787 (1995).
17. Osterwalder, J. *et al.* Angle-resolved photoemission experiments on $Bi_2Sr_2CaCu_2O_{8+\delta}$ (001): effects of the incommensurate lattice modulation. *Appl. Phys. A* **60,** 247–254 (1995).
18. Aebi, P. *et al.* Complete Fermi surface mapping of $Bi_2Sr_2CaCu_2O_{8+x}$(001): coexistence of short range antiferromagnetic correlations and metallicity in the same phase. *Phys. Rev. Lett.* **72,** 2757–2760 (1994).
19. Nakayama, K. *et al.* Shadow bands in single-layered $Bi_2Sr_2CuO_{6+\delta}$ studied by angle-resolved photoemission spectroscopy. *Phys. Rev. B* **74,** 054505 (2006).
20. Meng, J. Q. *et al.* Monotonic d-wave superconducting gap of the optimally doped $Bi_2Sr_{1.6}La_{0.4}CuO_6$ superconductor by laser-based angle-resolved photoemission spectroscopy. *Phys. Rev. B* **79,** 024514 (2009).
21. Meng, J. Q. *et al.* Growth, characterization and physical properties of high-quality large single crystals of $Bi_2(Sr_{2-x}La_x)CuO_{6+\delta}$ high-temperature superconductors. *Supercond. Sci. Technol.* **22,** 045010 (2009).
22. Ono S. & Ando, Y. Evolution of the resistivity anisotropy in $Bi_2Sr_{2-x}La_xCuO_{6+\delta}$ single crystals for a wide range of hole doping. *Phys. Rev. B* **67,** 104512 (2003).
23. Chakravarty, S. & Kee, H. Y. Fermi pockets and quantum oscillations of the Hall coefficient in high-temperature superconductors. *Proc. Natl Acad. Sci. USA* **105,** 8835–8839 (2008).
24. Chang, J. *et al.* Electronic structure near the 1/8-anomaly in La-based cuprates. *N. J. Phys.* **10,** 103016 (2008).
25. Chakravarty, S., Laughlin, R. B., Morr, D. K. & Nayak, C. Hidden order in the cuprates. *Phys. Rev. B* **63,** 094503 (2001).
26. Wen, X. G. & Lee, P. A. Theory of underdoped cuprates. *Phys. Rev. Lett.* **76,** 503–506 (1996).
27. Yang, K. Y., Rice, T. M. & Zhang, F. C. Phenomenological theory of the pseudogap state. *Phys. Rev. B* **73,** 174501 (2006).
28. Ng, T.-K. Spinon-holon binding in t-J model. *Phys. Rev. B* **71,** 172509 (2005).
29. Kaul, R. K., Kim, Y. B., Sachdev, S. & Senthil, T. Algebraic charge liquids. *Nature Phys.* **4,** 28–31 (2008).
30. Wise, W. D. *et al.* Charge-density-wave origin of cuprate checkerboard visualized by scanning tunnelling microscopy. *Nature Phys.* **4,** 696–699 (2008).
31. Ding, H. *et al.* Evolution of the Fermi surface with carrier concentration in $Bi_2Sr_2CaCu_2O_{8+\delta}$. *Phys. Rev. Lett.* **78,** 2628–2631 (1997).
32. Norman, M. R., Randeria, M., Ding, H. & Campuzano, J. C. Phenomenology of the low-energy spectral function in high-T_c superconductors. *Phys. Rev. B* **57,** R11093–R11096 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D.-H. Lee, P. A. Lee, S. Sachdev, Z.-X. Shen, X. G. Wen, Z. Y. Weng, T. Xiang, G. M. Zhang and F. C. Zhang for discussions. This work was supported by the NSFC, the MOST of China and the Chinese Academy of Sciences.

Author Contributions J.M. contributed to La-Bi2201 sample growth with the assistance of G.L.; X.D. and W.L. contributed to the magnetic measurement of samples; G.L., W.Z., L.Z., H.L., J.M., X.D., X.J., D.M., S.L., J.Z., G.W., Y. Zhou, Y. Zhu, X.W., Z.X. and C.C. contributed to the development and maintenance of the laser-ARPES system; J.M. carried out the experiment with assistance from G.L., W.Z., L.Z., H.L., X.J., D.M. and S.L.; X.J.Z. and J.M. analysed the data and wrote the paper; X.J.Z. was responsible for overall project direction, planning, management and infrastructure.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to X.J.Z. (XJZhou@aphy.iphy.ac.cn).

LETTERS

Ultraflat graphene

Chun Hung Lui[1], Li Liu[2], Kin Fai Mak[1], George W. Flynn[2] & Tony F. Heinz[1]

Graphene, a single atomic layer of carbon connected by sp^2 hybridized bonds, has attracted intense scientific interest since its recent discovery[1]. Much of the research on graphene has been directed towards exploration of its novel electronic properties, but the structural aspects of this model two-dimensional system are also of great interest and importance. In particular, microscopic corrugations have been observed on all suspended[2] and supported[3–8] graphene sheets studied so far. This rippling has been invoked to explain the thermodynamic stability of free-standing graphene sheets[9]. Many distinctive electronic[10–12] and chemical[13–15] properties of graphene have been attributed to the presence of ripples, which are also predicted to give rise to new physical phenomena[16–26] that would be absent in a planar two-dimensional material. Direct experimental study of such novel ripple physics has, however, been hindered by the lack of flat graphene layers. Here we demonstrate the fabrication of graphene monolayers that are flat down to the atomic level. These samples are produced by deposition on the atomically flat terraces of cleaved mica surfaces. The apparent height variation in the graphene layers observed by high-resolution atomic force microscopy (AFM) is less than 25 picometres, indicating the suppression of any existing intrinsic ripples in graphene. The availability of such ultraflat samples will permit rigorous testing of the impact of ripples on various physical and chemical properties of graphene.

The morphology of high-quality graphene crystals has been the subject of much attention. Detailed electron-diffraction studies of free-standing graphene monolayers[2] indicate the presence of an intrinsic rippling, with ~1-nm-high corrugations normal to the surface appearing over a characteristic lateral scale of 10–25 nm. It has been argued that these corrugations are necessary to stabilize the suspended graphene sheets against thermal instabilities present in ideal two-dimensional systems[9]. A comparable degree of height variation has also been reported in several studies of graphene monolayers deposited on insulating substrates[3–8]. This rippling has been invoked to explain many phenomena observed in graphene, such as the formation of electron–hole puddles[11,12], the suppression of weak localization[10], decreased carrier mobility[22] and enhanced chemical reactivity[13–15]. In addition, theoretical studies of graphene have predicted that graphene ripples will induce fascinating new phenomena, including the enhancement of spin–orbit coupling[16], the generation of an inhomogeneous density of states and the formation of zero-energy Landau levels in the absence of magnetic fields[17–20,23–25].

We report here the fabrication and characterization of high-quality ultraflat graphene monolayers by making use of a mica support that provides atomically flat terraces over large areas. Using high-resolution, non-contact mode atomic force microscopy (AFM) to characterize the morphology, we find that graphene on mica approaches the limit of atomic flatness. The apparent height variation of graphene on mica is found to be <25 pm over micrometre lateral length scales. This flatness, measured with a lateral spatial resolution of 7 nm, appears to be limited by instrument noise and is essentially identical (within 5 pm) to that observed for the surface of cleaved graphite crystals. Our results show that any intrinsic instability of graphene can be fully suppressed by deposition on an appropriate substrate. The availability of such a flat substance provides insight into questions of thermodynamic stability for this model two-dimensional system and also a reference material with which to determine the role of ripples in the panoply of observed and predicted phenomena.

The key to our experiments was the preparation of an atomically flat substrate for deposition of single-layer graphene crystals. For this purpose, we chose mica, a material composed of negatively charged aluminosilicate layers that are linked by single layers of potassium ions[27]. Since cleavage takes place readily along the potassium layer, atomically smooth surfaces with lateral dimensions as large as 100 μm can be routinely produced. Graphene monolayers were prepared by the standard method of mechanical exfoliation of kish graphite[1] on both mica and bulk SiO_2 substrates for comparative studies (see Methods).

We employed amplitude-modulation AFM in the non-contact mode to characterize the topography of the graphene samples. The AFM lateral and height resolution under scanning conditions were 7 nm and 23 pm, respectively. The AFM topographic images displayed in this paper are presented without filtering or smoothing. A third-order line and plane subtraction correction was applied to compensate for scanning drift and image bow. The roughness of the surface was characterized by the standard deviation σ of height distribution and the height correlation length l (see Supplementary Methods).

AFM topographic images acquired for regions surrounding the edges of graphene samples on both SiO_2 and mica substrates are shown in Fig. 1a, b. Histograms of the corresponding height distribution over the $200 \times 200\ nm^2$ regions of the surfaces are presented in Fig. 1c. For the bare SiO_2 surface, the parameters describing the height variation and correlation length (Table 1) are, respectively, $\sigma = 168$ pm and $l = 16$ nm. For the graphene monolayer on SiO_2, we find a comparable (or slightly diminished) degree of roughness, with $\sigma = 154$ pm and $l = 22$ nm, indicating that graphene monolayers largely follow the underlying substrate morphology.

In sharp contrast to these results, our AFM images on the mica substrate exhibit a much smoother landscape. For the bare mica surface, we obtain (see Table 1) $\sigma = 34.3$ pm and $l = 2$ nm. (As discussed below, we attribute the low value of l to residual AFM noise, rather than to physically meaningful features.) Taking the measured value of σ as a guide, the surface of mica is seen to be at least five times smoother than that of the SiO_2 substrate. When placed on such a flat mica terrace, graphene monolayers display an exceedingly flat structure, one quite different from that observed for graphene/SiO_2. This difference can be seen immediately by comparing the three-dimensional presentation of the AFM topographic images in Fig. 2a, b. More quantitatively, for graphene on mica, we obtain $\sigma = 24.1$ pm and $l = 2$ nm. This topography is at least five times smoother than that of graphene on SiO_2. Since the interlayer distance in bulk graphite is 340 pm, with an

[1]Departments of Physics and Electrical Engineering, Columbia University, 538 West 120th Street, [2]Department of Chemistry, Columbia University, 3000 Broadway, New York, New York 10027, USA.

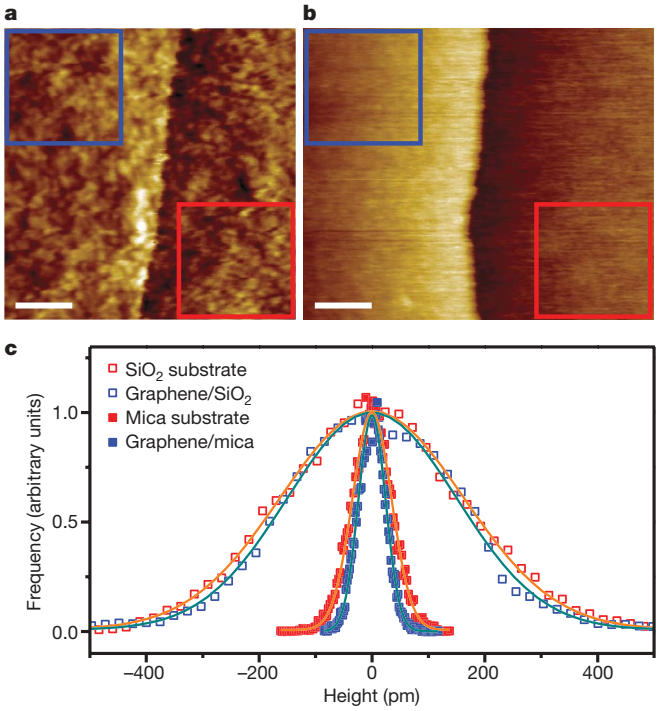


Figure 1 | AFM topographic images of different samples and the corresponding histograms of height. **a**, AFM image of a boundary between a graphene monolayer and a SiO_2 substrate. Graphene occupies the left-hand side of the image and the scale bar is 100 nm in length. **b**, As **a** for a graphene monolayer on a mica substrate. **c**, Height histograms for graphene on mica (solid blue squares), the mica substrate (solid red squares), graphene on SiO_2 (open blue squares), and the SiO_2 substrate (open red squares). The data, corresponding to the regions designated by the blue and red squares in the images of **a** and **b**, are described by Gaussian distributions (solid lines) with standard deviations σ of 24.1 pm, 34.3 pm, 154 pm, and 168 pm, respectively.

observed height variation of only 24.1 pm, we can consider graphene on mica as having reached the limit of atomic flatness with respect to ripples, that is, a height variation that is much less than the diameter of an atom when probed with our lateral resolution of 7 nm (see Supplementary Discussion).

Table 1 | σ and l of the images for different surfaces

	SiO_2	Graphene/SiO_2	Mica	Graphene/mica	Graphite
σ (pm)	168	154	34.3	24.1	22.6
l (nm)	16	22	2	2	2

The discussion of the flatness of the graphene/mica surface given above has been conservative in not attributing any of the observed height variation in the AFM images to instrumental noise. In fact, the results indicate that AFM noise is significant in measurements of flat surfaces. In particular, the correlation length of $l \approx 2$ nm calculated for the mica and the graphene/mica surfaces must arise largely from AFM noise, because any true physical features could only contribute to a correlation length comparable to or greater than the AFM spatial resolution of 7 nm. To address this issue, we made AFM measurements of the topography of cleaved kish graphite (Fig. 2c). The observed topography for the cleaved graphite surface is very similar to that of graphene/mica. Figure 2d compares the height histograms for graphite and graphene/mica. The widths of the distributions are, respectively, $\sigma = 22.6$ pm and $\sigma = 24.1$ pm. If we treat the graphite surface as entirely flat, then the measured standard deviation reflects the instrumental noise. Under the assumption that the instrumental noise adds in quadrature to any true height fluctuations, the values given above constrain the actual roughness of the graphene/mica sample to less than 8.5 pm.

Finally, in assessing the flatness of graphene, possible perturbations in its topography from tip-sample interactions must also be considered. To exclude the possibility that the observed flatness of graphene on mica might be produced through the suppression of ripples by the tip-sample interaction of the AFM probe, our AFM measurements have been carried out in a strictly non-contact mode, that is, one in which the tip/sample interaction was always attractive. Instead of pressing down on the surface at any time, the AFM tip is actually pulling slightly on the graphene sheet (see Supplementary Methods).

Since the discovery of intrinsic ripples in free-standing graphene, there has been considerable discussion of the role of substrate corrugation in determining the morphology of supported graphene

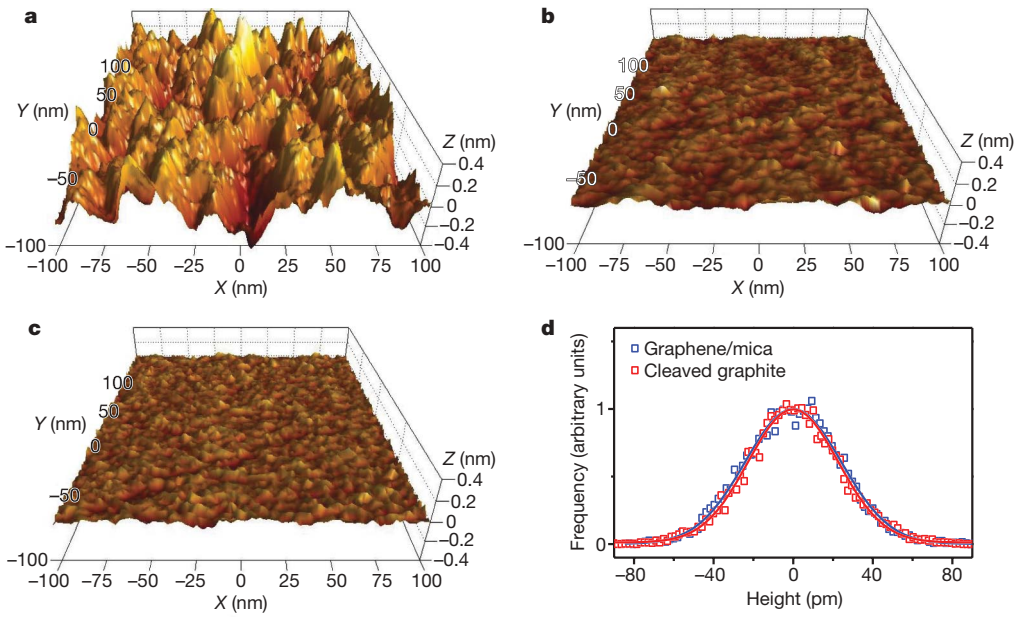


Figure 2 | Comparison of surface roughness for graphene on SiO_2 and on mica, and for cleaved graphite. **a**, **b**, Three-dimensional representations of the AFM topographic data for graphene on SiO_2 (**a**) and on mica (**b**) substrates. The images correspond to the regions in Fig. 1a and b designated by the blue squares. **c**, AFM image of the surface of a cleaved kish graphite sample. Images **a**, **b** and **c** correspond to 200 nm × 200 nm areas and are presented with the same height scale. **d**, Height histograms of the data in **b** as blue squares and in **c** as red squares. The histograms are described by Gaussian distributions (solid lines) with standard deviations σ of 24.1 pm and 22.6 pm, respectively.

monolayers[3–8]. Although the observed corrugation of supported graphene might well be an intrinsic feature[2,9] of the graphene monolayers in the experiments performed to date, a different explanation is equally possible. The roughness of the graphene surfaces may simply reflect the contours of the underlying substrates, which typically exhibit corrugation comparable to that observed in the supported graphene monolayers. Our measurements demonstrate unambiguously that intrinsic ripples in graphene, if they do exist, can be strongly suppressed by interfacial van der Waals interactions when this material is supported on an appropriate atomically flat substrate.

METHODS SUMMARY

Sample preparation. In our study we made use of grade V-1 muscovite mica substrates (15×15 mm^2, from Structure Probe) and produced graphene layers by the standard method of mechanical exfoliation of kish graphite[1]. Mica surfaces are known to be hydrophilic and readily adsorb water and carbon dioxide, as well as hydrocarbons. To minimize the presence of adsorbates at the graphene–mica interface, sample preparation was carried out in a glove box with water and oxygen concentrations below 1 p.p.m. (one part per million). For comparative studies, graphene monolayers were also prepared on bulk SiO_2 substrates. The SiO_2 substrates were carefully cleaned by sonication in methanol and the graphene samples were deposited by the same method of exfoliation of kish graphite, in this instance under ambient conditions. None of the samples described in this paper was subjected to any thermal processing.

Sample characterization. Graphene monolayers were identified on the mica substrate by optical microscopy, which was performed under ambient conditions. Although it is more difficult than for graphene samples deposited on an optimized SiO_2 overlayer on a silicon substrate, we were able to identify graphene monolayers directly by visual inspection. The modulation in reflectivity from a graphene monolayer still amounts to about 5% (Supplementary Fig. 1). Raman spectroscopy was applied for further characterization of the graphene samples[28] (Supplementary Fig. 2). From examination of the 2D mode Raman line, we confirmed the single-layer thickness of all the samples investigated in this paper. Also, the Raman spectra did not show any measurable D peak, indicating the high crystalline order of our samples. Our method of sample preparation was found to produce a significant yield of large graphene monolayers, with characteristic lateral dimensions ranging from tens of micrometres up to ~0.2 mm. The efficient deposition of large graphene single layers is attributed to the flat and clean surface of freshly cleaved mica.

Received 20 July 2007; accepted 1 October 2009.

1. Novoselov, K. S. *et al.* Two-dimensional atomic crystals. *Proc. Natl Acad. Sci. USA* **102,** 10451–10453 (2005).
2. Meyer, J. C. *et al.* The structure of suspended graphene sheets. *Nature* **446,** 60–63 (2007).
3. Stolyarova, E. *et al.* High-resolution scanning tunneling microscopy imaging of mesoscopic graphene sheets on an insulating surface. *Proc. Natl Acad. Sci. USA* **104,** 9209–9212 (2007).
4. Ishigami, M., Chen, J. H., Cullen, W. G., Fuhrer, M. S. & Williams, E. D. Atomic structure of graphene on SiO_2. *Nano Lett.* **7,** 1643–1648 (2007).
5. Booth, T. J. *et al.* Macroscopic graphene membranes and their extraordinary stiffness. *Nano Lett.* **8,** 2442–2446 (2008).
6. Knox, K. R. *et al.* Spectromicroscopy of single and multilayer graphene supported by a weakly interacting substrate. *Phys. Rev. B* **78,** 201408 (2008).
7. Stoberl, U., Wurstbauer, U., Wegscheider, W., Weiss, D. & Eroms, J. Morphology and flexibility of graphene and few-layer graphene on various substrates. *Appl. Phys. Lett.* **93,** 051906 (2008).
8. Geringer, V. *et al.* Intrinsic and extrinsic corrugation of monolayer graphene deposited on SiO_2. *Phys. Rev. Lett.* **102,** 076102 (2009).
9. Fasolino, A., Los, J. H. & Katsnelson, M. I. Intrinsic ripples in graphene. *Nature Mater.* **6,** 858–861 (2007).
10. Morozov, S. V. *et al.* Strong suppression of weak localization in graphene. *Phys. Rev. Lett.* **97,** 016801 (2006).
11. Martin, J. *et al.* Observation of electron-hole puddles in graphene using a scanning single-electron transistor. *Nature Phys.* **4,** 144–148 (2008).
12. Deshpande, A., Bao, W., Miao, F., Lau, C. N. & LeRoy, B. J. Spatially resolved spectroscopy of monolayer graphene on SiO_2. *Phys. Rev. B* **79,** 205411 (2009).
13. Liu, L. *et al.* Graphene oxidation: thickness-dependent etching and strong chemical doping. *Nano Lett.* **8,** 1965–1970 (2008).
14. Ryu, S. *et al.* Reversible basal plane hydrogenation of graphene. *Nano Lett.* **8,** 4597–4602 (2008).
15. Elias, D. C. *et al.* Control of graphene's properties by reversible hydrogenation: evidence for graphane. *Science* **323,** 610–613 (2009).
16. Huertas-Hernando, D., Guinea, F. & Brataas, A. Spin-orbit coupling in curved graphene, fullerenes, nanotubes, and nanotube caps. *Phys. Rev. B* **74,** 155426 (2006).
17. de Juan, F., Cortijo, A. & Vozmediano, M. A. H. Charge inhomogeneities due to smooth ripples in graphene sheets. *Phys. Rev. B* **76,** 165409 (2007).
18. Brey, L. & Palacios, J. J. Exchange-induced charge inhomogeneities in rippled neutral graphene. *Phys. Rev. B* **77,** 041403 (2008).
19. Guinea, F., Horovitz, B. & Le Doussal, P. Gauge field induced by ripples in graphene. *Phys. Rev. B* **77,** 205421 (2008).
20. Guinea, F., Katsnelson, M. I. & Vozmediano, M. A. H. Midgap states and charge inhomogeneities in corrugated graphene. *Phys. Rev. B* **77,** 075422 (2008).
21. Herbut, I. F., Juricic, V. & Vafek, O. Coulomb interaction, ripples, and the minimal conductivity of graphene. *Phys. Rev. Lett.* **100,** 046403 (2008).
22. Katsnelson, M. I. & Geim, A. K. Electron scattering on microscopic corrugations in graphene. *Phil. Trans. R. Soc. A* **366,** 195–204 (2008).
23. Kim, E. A. & Neto, A. H. C. Graphene as an electronic membrane. *Europhys. Lett.* **84,** 57007 (2008).
24. Vozmediano, M. A. H., de Juan, F. & Cortijo, A. Gauge fields and curvature in graphene. *J. Phys. Conf. Ser.* **129,** 012001 (2008).
25. Wehling, T. O., Balatsky, A. V., Tsvelik, A. M., Katsnelson, M. I. & Lichtenstein, A. I. Midgap states in corrugated graphene: ab initio calculations and effective field theory. *Europhys. Lett.* **84,** 17003 (2008).
26. Cortijo, A. & Vozmediano, M. A. H. Minimal conductivity of rippled graphene with topological disorder. *Phys. Rev. B* **79,** 184205 (2009).
27. Bragg, S. L. & Claringbull, G. F. *Crystal Structures of Minerals* Ch. 13 (G. Bell and Sons, 1965).
28. Ferrari, A. C. *et al.* Raman spectrum of graphene and graphene layers. *Phys. Rev. Lett.* **97,** 187401 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E. B. Newton and S. Li for their assistance in sample preparation and J. Shan, H. G. Yan and Z. Q. Li for discussions. This work was supported by grants from DARPA through the CERA programme (to T.F.H.), from the Nano Electronics Research Corporation (NERC) through the INDEX Center (to T.F.H.), and from the National Science Foundation (grant CHE-07-01483 to G.W.F.). Equipment and material support was provided by the US Department of Energy (grant DE-FG02-88-ER13937 to G.W.F.).

Author Contributions All of the authors contributed to the design of the experiment; C.H.L. and K.F.M. were responsible for the sample preparation, C.H.L. and L.L. characterized the samples by AFM imaging; C.H.L., L.L., and T.F.H. devised the method for and performed the data analysis; and C.H.L., G.W.F. and T.F.H. prepared the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.F.H. (tony.heinz@columbia.edu).

LETTERS

Evidence for warmer interglacials in East Antarctic ice cores

L. C. Sime[1], E. W. Wolff[1], K. I. C. Oliver[2]† & J. C. Tindall[3]

Stable isotope ratios of oxygen and hydrogen in the Antarctic ice core record have revolutionized our understanding of Pleistocene climate variations and have allowed reconstructions of Antarctic temperature over the past 800,000 years (800 kyr; refs 1, 2). The relationship between the D/H ratio of mean annual precipitation and mean annual surface air temperature is said to be uniform ±10% over East Antarctica[3] and constant with time ±20% (refs 3–5). In the absence of strong independent temperature proxy evidence allowing us to calibrate individual ice cores, prior general circulation model (GCM) studies have supported the assumption of constant uniform conversion for climates cooler than that of the present day[3,5]. Here we analyse the three available 340 kyr East Antarctic ice core records alongside input from GCM modelling. We show that for warmer interglacial periods the relationship between temperature and the isotopic signature varies among ice core sites, and that therefore the conversions must be nonlinear for at least some sites. Model results indicate that the isotopic composition of East Antarctic ice is less sensitive to temperature changes during warmer climates. We conclude that previous temperature estimates from interglacial climates are likely to be too low. The available evidence is consistent with a peak Antarctic interglacial temperature that was at least 6 K higher than that of the present day —approximately double the widely quoted 3 ± 1.5 K (refs 5, 6).

Studies have shown that, in simple cases, the stable isotope composition of oxygen and hydrogen (δ; see Methods) in precipitation is controlled by the temperature difference between the evaporation site and the condensation site, owing to temperature dependent fractionation processes[7]. The δ value of ice in high latitude ice cores is therefore used as a proxy for past temperatures[7,8]. Past core site surface temperature (T) is obtained from δ by applying an estimated 'palaeothermometer' gradient a:

$$\frac{\partial(\delta)}{\partial T} = a \tag{1}$$

Temporally constant and almost geographically invariant East Antarctic gradients have previously been applied in conversions of δ into T. Dome C, Vostok and Dome F ice core δ was converted to temperature using a palaeothermometer gradient of 6‰ per K (ref. 5) or 6.4‰ per K (ref. 4), for δD. These conversions imply that past interglacial T peaked at about 3 ± 1.5 K above the present day temperature at these core sites[5,6].

The past 340 kyr of the δ record from the three deep East Antarctic icecores[5,6,9] are shown on the EDC3 (EPICA Dome C) timescale[10] in Fig. 1. Similarities to the deep-sea isotope record[11] suggest that these δ timeseries record aspects of global climate. However, individual ice core δ values diverge significantly in time. Dome F attains maximum peak interglacial values 20–140% higher than Dome C referenced to the present. For the 340 kyr records, the most widely geographically separated pair of Dome C and Dome F ice core δ values show the largest differences.

To explore the reasons for these differences, we calculate the temporal gradients in δ at each site. We then compare the gradient at each location (x, y) to that at Dome C:

$$R_\delta(x, y, t) = \frac{\dot{\delta}(x, y, t)}{\dot{\delta}(x_0, y_0, t)} = R_T(x, y, t) R_a(x, y, t)$$
$$R_T(x, y, t) = \frac{\dot{T}(x, y, t)}{\dot{T}(x_0, y_0, t)}, \quad R_a(x, y, t) = \frac{a(x, y, t)}{a(x_0, y_0, t)} \tag{2}$$

where (x_0, y_0) is the location of Dome C and overdots (for example, $\dot{\delta}$) denote differentiation with respect to time. This shows that R_δ is the product of two components: the first (R_T) due to geographical variability in the rate of change of temperature, and the second (R_a) due to geographical variability in the palaeothermometer gradient a. When R_δ is one, there is no geographical variation in $\dot{\delta}$ in the East Antarctic; R_δ above (below) one indicates that the ice core site $\dot{\delta}$ is higher (lower) than the Dome C record.

Cold climate R_δ median values of around 0.8 (Fig. 2, left side) indicate that a δ change of 1‰ at Dome C tends to coincide with a

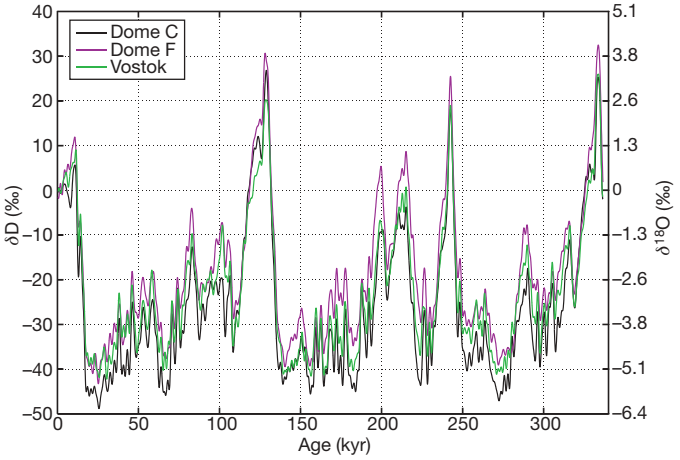


Figure 1 | Time series of Dome C, Dome F and Vostok ice core δ records from preindustrial times until 336.5 kyr ago. All time series are on the EDC3 timescale[10], on a 0.1 kyr interval. The records are presented as anomalies from the present day. See Methods for details of δ.

[1]British Antarctic Survey, Cambridge CB3 0ET, UK. [2]Department of Earth and Environmental Sciences, The Open University, Milton Keynes MK 7 6AA, UK. [3]School of Geographical Sciences, University of Bristol, University Road, Bristol BS8 1SS, UK. †Present address: School of Ocean and Earth Science, National Oceanography Centre, University of Southampton, Southampton SO14 3ZH, UK.

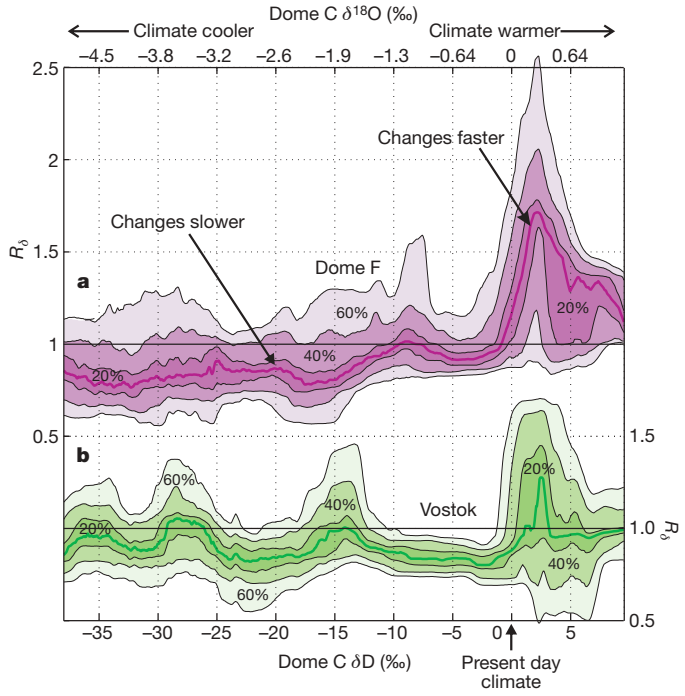


Figure 2 | Observed ice core R_δ against δ. Here R_δ is the ratio of $\dot{\delta}$ at Dome F (**a**) or at Vostok (**b**) relative to that at Dome C; the upper and lower x axes show respectively Dome C δ^{18}O and Dome C δD. The shading around each R_δ value shows the 20%, 40% and 60% envelope of the R_δ observations. We do not show R_δ for $\delta D < -38‰$ or $>8‰$ because of insufficient observations of R_δ. R_δ is binned according to Dome C δ with an approximately uniform δ bin width for each sample (see Methods).

change of 0.8‰ at Dome F and Vostok. However, during warmer interglacial periods (Fig. 2, right side), Dome F δ exhibits much stronger variation than Dome C, indicated by R_δ values above 1. Vostok R_δ has a qualitatively similar but much weaker warm climate R_δ peak. During past warm periods, with δD isotopic values 0–7‰ above present day values, a change of 1‰ at Dome C tends to coincide with a change of 1.4–1.7‰ at Dome F. However, from the observed R_δ ratio we cannot a priori deduce to what extent spatial variability in temperature change, R_T, or in the palaeothermometer gradient, R_a, determine R_δ.

Isotope enabled GCM palaeoclimate experiments provide a powerful means for unravelling the causes of geographical variations in $\dot{\delta}$ across East Antarctica. The isotopic GCM[12] we use has a good representation of the present day global[13] and Antarctic climate[14], and isotope distribution[12,15] (see Supplementary Information). Modelled values for East Antarctic R_δ, R_T and R_a are obtained using experiments representing the present day[16], Last Glacial Maximum (LGM)[17], and climates between 0.5 K and 3 K warmer than the modelled present day Dome C temperatures. Because current GCM experiments, using insolation and greenhouse-gas levels appropriate for the last interglacial (125–130 kyr ago), fail to reproduce warmer Southern Hemisphere[18] and Antarctic temperatures[19], we drive our warmer climate experiments instead with higher greenhouse-gas levels.

GCM experiment results in Fig. 3 indicate that geographical variability in temperature change R_T is small compared to geographical variability in the palaeothermometer R_a for both cooler and warmer climates. Additional GCM results[20] and independent calculations of R_T using GCM output confirm that R_T is small (see Supplementary Information) compared to observed R_δ. The spatially invariant rates of temperature change shown also suggest that relative accumulation changes between the sites are likely to be small, implying that relative elevation changes between the sites are also small. Analysis of our experiments support this: Dome F precipitation for the present day and all warmer experiments remains constantly 17 ± 4 m kyr^{-1} water equivalent lower than Dome C. Therefore we would expect relative elevation changes between the sites for the modelled warmed climates should be within ± 4 m kyr^{-1}. From these results we conclude that

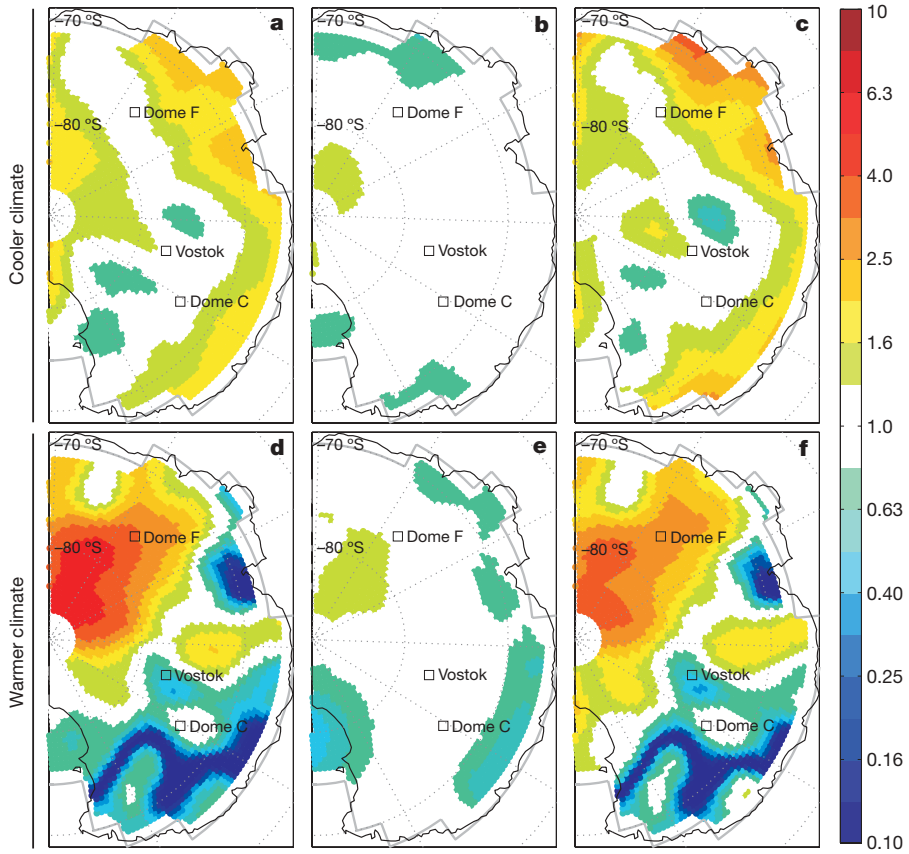


Figure 3 | The geographical pattern of $\dot{\delta}$, separated into a temperature change component R_T and a palaeothermometer component R_a. **a**, Mean R_δ (colour coded) between the present day and LGM climate (this is a discrete value calculated from the change between the modelled climates). **b**, **c**, As **a** but showing R_T and R_a respectively. We note that $\log R_\delta = \log R_T + \log R_A$. **d**–**f**, As **a**–**c** but for a change between the present day and a 3 K warmer (at Dome C) climate. All geographical ratios are referenced to the Dome C core site, hence Dome C R values are all 1. Results are shown using a logarithmic colour scale, and are calculated using values of δ and T with a 300 km radius averaging[15].

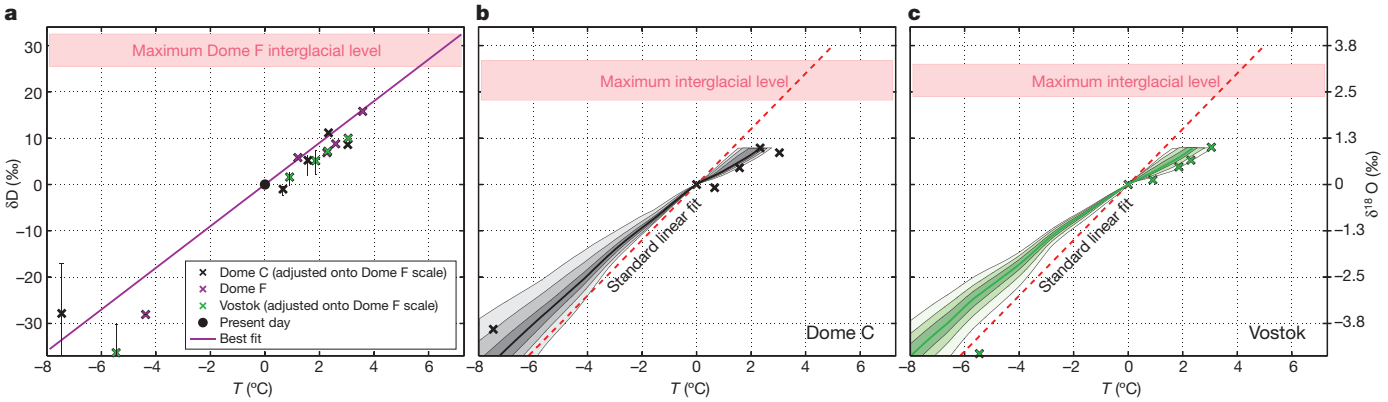


Figure 4 | The δ versus T palaeothermometer relationship. **a**, Dome F model results (magenta), and model results for the other sites normalized to Dome F by the median measured R_δ for the appropriate temperature. The line is the best fit through all the data, constrained to the zero point (see Methods). Error bars show the outer envelope (envelopes are shown in Fig. 2) for each normalization. **b**, **c**, Actual model results (crosses) for Dome C (**b**) and Vostok (**c**), and the best fit in **a** adjusted according to the observed R_δ. The median R_δ and envelope used are shown in Fig. 2. In **b** and **c**, the red dashed line shows the standard[4,5] conversion line $a = 6\delta$D, where a is in units of ‰ per K. The range of maximum interglacial core site δ values are indicated on each panel.

high plateau East Antarctic R_T is small and thus, for East Antarctic ice cores, R_a is the dominant component of R_δ.

The finding that R_δ approximately equals R_a suggests that the observed differences in the ice core δ records (Fig. 1) can be explained primarily by variations in the palaeothermometer gradient, rather than in temperature. Therefore, the standard approach of applying a constant palaeothermometer gradient must introduce errors to temperature reconstructions for at least two of the three core sites. Although probably too simple, the least complex assumption that is consistent with the data are that one site has a constant gradient, and can be treated as a reference site. We then normalize model δ values for the other sites to those of the reference, using integrated observed R_δ values for the appropriate temperature shown Fig. 2; this allows us to use the model results from all three sites together to make a best estimate of temperature changes (see Methods Summary). Using this approach, we find that the best least squares fit is obtained if the Dome F site has a constant gradient; this gives a palaeothermometer gradient of 4.5 ± 1.1‰ per K (Fig. 4). This gradient and the uncertainties calculated (see Methods) imply that maximum interglacial temperatures over the past 340 kyr were between 6.0 K and 10.0 K above present-day values (Fig. 4a).

Detailed analysis of GCM experiments shows trends in covariance between surface temperature and precipitation throughout the modelled warming[15], which can affect ice core δ records[7]. The seasonal and synoptic covariance changes have a limited impact on the geographical warming pattern R_T (refs 15, 21), but are sufficient to explain our climate dependence of the δ against T relationship. Warmer climates are associated with a larger proportion of precipitation in cold seasons over Dome C and Vostok[15]. This effectively reduces the condensation temperature change relative to the mean annual temperature change, lowering the palaeothermometer gradient at these sites for warm climates compared to cold climates. Over the Dome F sector, trends in high pressure blocking events have been observed over the past two decades[21,22]. Similarly, modelled synoptic trends in covariance occur in the GCM experiments over the Dome F sector[15] and act to raise the warm climate palaeothermometer gradient in this region. These changes in covariance make the Dome F δ response to T more weakly nonlinear, thus the constant Dome F conversion of δ to T is a better approximation than a constant conversion for Dome C or Vostok.

The majority of the present day precipitation falling across the East Antarctic Plateau, which accumulates to form ice cores, is evaporated from the Atlantic and Indian sectors of the Southern Ocean[23,24]. Changes in the evaporative source characteristics are thought to be relatively small for our core sites during climate shifts[25,26]. Additionally, because the source vapour is from a common region, changes in source temperature or humidity affect East Antarctic Plateau precipitation uniformly during climate changes. For this reason evaporative source changes do not qualitatively affect the analysis above: applying a constant palaeothermometer gradient introduces errors in two of the three sites, regardless of any source temperature or humidity changes. Most uncertainty in the relationship between warm period temperature and δ is due to changes in the Antarctic accumulation regime[15], described in the previous paragraph, and complex atmospheric mixing changes such as those induced by warm climate differences in sea-ice and rates of atmospheric overturning[27,28]. Analysis of additional warm period GCM experiments, which feature different warm climate sea-ice, sea surface temperatures, and precipitation regimes, indicate that these warm period climate differences contribute ± 1.1‰ per K to the gradient (see Online Methods).

An interglacial temperature of the level suggested by Fig. 4 indicates that there are serious deficiencies in our understanding of warmer than present day climates[18,19]. It may be that ice sheet elevation changes outside the East Antarctic, such as the possible loss of the West Antarctic ice-sheet[29], or changes in sea surface temperatures and sea-ice, have contributed to the past high interglacial East Antarctic temperatures. Coupled ocean–atmosphere GCM experiments could throw light on the causes of these high East Antarctic temperatures. The analysis presented here can be used to assess the results of such modelling studies. However, further proxy data on all aspects of interglacial climates would be invaluable. Past interglacials offer the chance to test whether modelled consequences of future polar warming are realistic. It is therefore essential to confirm what temperature level was reached, before one can assess the consequences of such a climate.

METHODS SUMMARY

East Antarctic stable water isotope (deuterium, D, and oxygen-18, ^{18}O) records from 336.5 kyr ago are available for three ice cores[5,6,9]. Fractional isotopic content is expressed for deuterium as: $\delta\mathrm{D} = 1{,}000 \times [(\mathrm{HD}^{16}\mathrm{O}/\mathrm{H_2}^{16}\mathrm{O})/R_{\mathrm{VSMOW}} - 1]$ (in ‰), where R_{VSMOW} is the ratio of $\mathrm{HD}^{16}\mathrm{O}$ to $\mathrm{H_2}^{16}\mathrm{O}$ for Vienna standard mean ocean water[8]. We use δ as shorthand to represent both δD and δ^{18}O (see Online Methods), put all values on the EDC3 timescale[10], and present δ values as anomalies from the present day. Because of differences in the temporal resolution of the records, a 1.5 kyr low-pass filter is used on the δ records to ensure they are equivalent. We use 3 kyr segments to calculate a discrete approximation to $\dot{\delta}$, and exclude $\dot{\delta}$ values which show sensitivity to ice core synchronization errors. The $\dot{\delta}(t)$ values are used to calculate $R_\delta(t)$ (equation (2)). Values are then binned according to Dome C $\delta(t)$ to get $R_\delta(\delta)$ (Fig. 2). We use larger sample sizes for colder climates to provide an approximately uniform δ bin width.

A best fit through the GCM experiments (see Supplementary Information), constrained to the zero point, is obtaining by calculating the value of a_0 that minimizes $\sum E_{ij}^2$ in

$$\delta_{ij}\int_0^{T_{ij}}(R_\delta)_{j\mathrm{ref}}\mathrm{d}T=\left(a_0T_{ij}+E_{ij}\right)\int_0^{T_{ij}}(R_\delta)_j\mathrm{d}T$$

where T_{ij} and δ_{ij} are GCM temperature and δ in simulation i at core site j, $j$ref is the choice of reference site where a constant palaeothermometer gradient is assumed, and R_δ at Dome C is 1. The optimal r.m.s. (root-mean-squared) errors using Dome C, Dome F and Vostok as reference sites are 6.0‰, 4.3‰ and 5.1‰, respectively. These r.m.s. values indicate that it is more appropriate to assume a linear relation for Dome F than for the other sites. The method for estimating the uncertainty in a_0 is described in the Online Methods section.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 October 2008; accepted 5 October 2009.

1. Solomon, S. *et al.* (eds) *Climate Change 2007: The Physical Science Basis* (Cambridge Univ. Press, 2007).
2. EPICA Community Members. Eight glacial cycles from an Antarctic ice core. *Nature* **429,** 623–628 (2004).
3. Jouzel, J. *et al.* Magnitude of isotope/temperature scaling for interpretation of central Antarctic ice cores. *J. Geophys. Res.* **108** (D12), 26471–26487 (2003).
4. Watanabe, O. *et al.* Homogeneous climate variability across East Antarctica over the past three glacial cycles. *Nature* **422,** 509–512 (2003).
5. Jouzel, J. *et al.* Orbital and millennial Antarctic climate variability over the past 800,000 years. *Science* **317,** 793–796 (2007).
6. Petit, J. R. *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399,** 429–436 (1999).
7. Dansgaard, W. Stable isotopes in precipitation. *Tellus* **16,** 436–468 (1964).
8. Rozanski, K., Araguas-Araguas, L. & Gonfiantini, R. Relation between long-term trends of oxygen-18 isotope composition of precipitation and climate. *Science* **258,** 981–985 (1992).
9. Kawamura, K. *et al.* Northern Hemisphere forcing of climatic cycles in Antarctica over the past 360,000 years. *Nature* **448,** 912–916 (2007).
10. Parrenin, F. *et al.* The EDC3 chronology for the EPICA Dome C ice core. *Clim. Past* **3,** 485–497 (2007).
11. Lisiecki, L. E. & Raymo, M. E. A Pliocene-Pleistocene stack of 57 globally distributed benthic δ^{18}O records. *Paleoceanography* 20, PA1003, doi:10.1029/2004PA001071 (2005).
12. Tindall, J., Valdes, P. & Sime, L. Stable water isotopes in HadCM3: the isotopic signature of ENSO and the tropical amount effect. *J. Geophys. Res.* D04111, doi:10.1029/2008JD010825 (2008).
13. Pope, V. D., Gallani, M. L., Rowntree, P. R. & Stratton, R. A. The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3. *Clim. Dyn.* **16,** 123–146 (2000).
14. Turner, J., Connolley, W. M., Lachlan-Cope, T. A. & Marshall, G. J. The performance of the Hadley Centre Climate Model (HadCM3) in high southern latitudes. *Int. J. Climatol.* **26,** 91–112 (2006).
15. Sime, L. C., Tindall, J., Wolff, E., Connolley, W. & Valdes, P. The Antarctic isotopic thermometer during a CO_2 forced warming event. *J. Geophys. Res.* **113,** D24119, doi:10.1029/2008JD010395 (2008).
16. Rayner, N. A. *et al.* Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.* **108** (D14), 4407, doi:10.1029/2002JD002670 (2003).
17. Paul, A. & Schäfer-Neth, C. Modeling the water masses of the Atlantic Ocean at the Last Glacial Maximum. *Paleoceanography* **18,** 1058, doi:10.1029/2002PA000783 (2003).
18. Otto-Bliesner, B. L. *et al.* Simulating Arctic climate warmth and icefield retreat in the last interglaciation. *Science* **311,** 1751–1753 (2006).
19. Groll, N., Widmann, M., Jones, J., Kaspar, F. & Lorenz, S. Simulated differences in the relationships between regional temperatures and large-scale circulation during the early Eemian interglacial (125 kyr BP) and the pre-industrial period. *J. Clim.* **18,** 4035–4048 (2005).
20. Bracegirdle, T. J., Connolley, W. M. & Turner, J. Antarctic climate change over the twenty first century. *J. Geophys. Res.* **113,** D03103, doi:10.1029/2007JD008933 (2008).
21. Schneider, D. P., Steig, E. J. & Comiso, J. C. Recent climate variability in Antarctica from satellite-derived temperature data. *J. Clim.* **17,** 1569–1583 (2004).
22. Hirasawa, N., Nakamura, H. & Yamanouchi, T. Abrupt changes in meteorological conditions observed at an inland Antarctic station in association with wintertime blocking. *Geophys. Res. Lett.* **27,** 1911–1914 (2000).
23. Delaygue, G., Jouzel, J., Masson, V., Koster, R. D. & Bard, E. Validity of the isotopic thermometer in central Antarctica: limited impact of glacial precipitation seasonality and moisture origin. *Geophys. Res. Lett.* **27,** 2677–2680 (2000).
24. Werner, M., Heimann, M. & Hoffmann, G. Isotopic composition and origin of polar precipitation in present and glacial climate simulations. *Tellus B* **53,** 53–71 (2001).
25. Vimeux, F., Masson, V., Jouzel, J., Stievenard, M. & Petit, J. R. Glacial–interglacial changes in ocean surface conditions in the Southern Hemisphere. *Nature* **398,** 410–413 (1999).
26. Vimeux, F., Cuffey, K. & Jouzel, J. New insights into Southern Hemisphere temperature changes from Vostok ice cores using deuterium excess correction. *Earth Planet. Sci. Lett.* **203,** 829–843 (2002).
27. Noone, D. & Simmonds, I. Sea ice control of water isotope transport to Antarctica and implications for ice core interpretation. *J. Geophys. Res.* **109,** D07105, doi:10.1029/2003JD004228 (2004).
28. Noone, D. The influence of midlatitude and tropical overturning circulation on the isotopic composition of atmospheric water vapor and Antarctic precipitation. *J. Geophys. Res.* **113,** D04102, doi:10.1029/2007JD008892 (2008).
29. Overpeck, J. T. *et al.* Paleoclimatic evidence for future ice-sheet instability and rapid sea-level rise. *Science* **311,** 1747–1750 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank W. Connolley for assistance with model set-up; T. Bracegirdle for organizing multi-model AR4 output; NERC RAPID ISOMAP for funding the model development; and the modelling groups, and the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model data set.

Author Contributions L.C.S. and E.W.W. discussed the original concept for the work. L.C.S. and K.I.C.O. wrote the ice core R_δ analysis. J.C.T. wrote the isotopic code for the HadAM3 model. L.C.S. set up and analysed the isotopic HadAM3 experiments and analysed the additional AR4 GCM output. L.C.S. wrote the paper. All authors discussed the results and modified the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to L.C.S. (lsim@bas.ac.uk).

METHODS

East Antarctic Dome C and Vostok deuterium (D) and Dome F oxygen-18 (^{18}O) records from the preindustrial until 336.5 kyr ago are available at the World Data Center for Paleoclimatology[5,6,9]. The fractional contents of the stable water isotopes D and ^{18}O are expressed as deviations from a standard water isotope sample, so for deuterium: $\delta D = 1{,}000 \times [(HD^{16}O/H_2{}^{16}O)/R_{VSMOW} - 1]$ (units are ‰), and for ^{18}O: $\delta^{18}O = 1{,}000 \times \left[\left(H_2^{18}O/H_2{}^{16}O\right)/R_{VSMOW} - 1\right]$ (units are ‰), with R_{VSMOW} = the ratio of Vienna standard mean ocean water for each isotopic species[8]. Here we use δ as shorthand to represent both δD and δ^{18}O.

The timeseries of Dome C, Dome F and Vostok ice core δ records are placed on the EDC3 timescale[10]. The timeseries are put on an 0.1 kyr interval. Records are smoothed with a 0.2 kyr running mean to prevent any high frequency aliasing. They are then aligned on the *y*-axis using the mean of the last 3 kyr of each record, that is, the present day δ of each record, thus δ values are presented as anomalies from the present day. A 1.5 kyr low-pass filtering ensures that changes in the temporal resolution of the records (which drops to 0.65 kyr for older parts of the Dome F record) do not affect the sampling throughout and between the records.

The relationship between $\delta^{18}O(t)$ and $\delta D(t)$ is generally presented in terms of a conversion coefficient plus a small temporally varying term, the deuterium excess or $d(t)$ parameter. The conversion is defined here as $\delta D(t) = d(t) + 7.85 \times \delta^{18}O(t)$, with no uncertainties, using the East Antarctic coefficient of 7.85 (ref. 25) from the Vostok ice core. The Dome F $d(t)$ record[30] can therefore be used to convert the Dome F $\delta^{18}O(t)$ record into a $\delta D(t)$ record so that $\delta(t)$ can be directly compared for the three cores. We also use unpublished Dome C $\delta^{18}O(t)$ and Dome F $\delta D(t)$ to ensure that no extra errors are introduced into the analysis. The standard deviation of Dome C $d(t)$ is 1.7‰, and of this, 21% of $d(t)$ correlates with $\delta(t)$. Because the standard deviation (a measure of the variance) is very small, we use the mean coefficient of 7.85 to provide results on dual δD and δ^{18}O axes. Presenting all δD and δ^{18}O values relative to the present day ensures that errors due to using dual axes are minimized. Of this small uncertainty introduced to the axes by $d(t)$, 79% is uncorrelated with $\delta(t)$, so systematic errors associated with the dual axes are negligible compared with the δ ranges considered in this manuscript. Note, as described above, that this negligible dual axes error does not affect the calculated R_δ values.

We use 3 kyr segments of $\delta(t)$ to calculate a discrete approximation to $\dot{\delta}(t)$ for each site, that is, the change in δ through time. The $\dot{\delta}(t)$ values are used to calculate $R_\delta(t)$ (equation (2)), and $R_\delta(t)$ is then binned according to Dome C δ (Fig. 2) to obtain $R_\delta(\delta)$. Because of the high frequency of low $\delta(t)$ values, and the low frequency of high $\delta(t)$ values, sample sizes are larger for colder climates (sample size 20 kyr at $\delta D = -37$‰), and linearly decreases for less depleted (warmer) climates (sample size 4 kyr at $\delta D = 9$‰). This gives an approximately uniform bin width.

Because the R_δ analysis of the ice core $\delta(t)$ values is based on ratios of $\dot{\delta}(t)$ (temporal change in δ), R_δ is almost completely independent of absolute age-model accuracy (as R_δ is a geographically relative rather than an absolute measure). Similarly, marine source changes due to ice-sheet volume changes[26] do not affect the calculation of R_δ, because all records are affected identically. Likewise, marine source temperature and humidity changes[25,26] will have little effect on R_δ because they vary little across the plateau[23,24]. The R_δ analysis will however be sensitive to ice core synchronization errors. We calculate the variability in $\delta(t)$ which can be explained by a linear trend in each 3 kyr segment, and do not use 3 kyr segments where more than 50% of the variability in $\delta(t)$ in either record is not explained by a linear fit. This limits the analysis to portions of the ice core where both records show a simple linear trend, eliminating very low $\dot{\delta}(t)$ values and times of inflecting $\delta(t)$ values (uncertain $\dot{\delta}(t)$). Removing these low and uncertain $\dot{\delta}(t)$ periods from the analysis minimizes the impact of uncertainties in the depth–age model on the calculated R_δ. The use of 3-kyr segments means we are using a discrete approximation to $\dot{\delta}(t)$, which would be unsuitable for examining variability on seasonal to millennial timescales.

We calculate the best fit through all GCM experiments (see Supplementary Information for details of model and boundary conditions), constrained to the zero point, by obtaining the value of a that minimizes $\sum E_{ij}^2$ in

$$\left(\delta_{ij} + \hat{\delta}_{ij}\right)\left(I_{j\mathrm{ref}} + \hat{I}_{j\mathrm{ref}}\right) = \left(aT_{ij} + E_{ij}\right)\left(I_j + \hat{I}_j\right), \quad I_j = \int_0^{T_{ij}} (R_\delta)_j \mathrm{d}T$$

where T_{ij} and δ_{ij} are GCM temperature anomalies and δ in simulation i at core site j, $j$ref is the choice of reference site where a constant palaeothermometer gradient is assumed, and R_δ at Dome C is 1. $\hat{\delta}$ and $\hat{I}$ are the error in δ and I, and are assumed equal to zero in determining the best estimate gradient a_0 of 4.5‰ K^{-1} with Dome F as the reference site. The optimal r.m.s. values of E_{ij} using Dome C, Dome F and Vostok as reference sites are 6.0‰, 4.3‰ and 5.1‰, respectively. These r.m.s. error values indicate that it is more appropriate to assume a linear relation for Dome F than for the other sites.

Uncertainties associated with this method for estimating the palaeothermometer gradient arise principally by two mechanisms: (1) uncertainty in observational R_δ (see envelopes in Fig. 2), which relates both to observed 'climate noise' and measurement (for example, age-model) noise; and (2) uncertainty in the model palaeothermometer, whereby different climates yielding identical temperatures at a given site may yield different values of δ. These uncertainties are represented in the error terms $\hat{I}$ and $\hat{\delta}$, respectively. A 1,000-member Monte Carlo analysis is used to obtain a distribution of palaeothermometer gradients resulting from applying values of $\hat{I}_{ij}$ and $\hat{\delta}_{ij}$ randomly selected from Gaussian distributions. The distribution for $\hat{I}$ is determined using median R_δ values (presented in Fig. 2), and observed δ at Dome F, to estimate δ at Dome C and Vostok. The r.m.s. error in this estimate of δ, compared to observed δ, is 4.4‰ at Dome C and 3.5‰ at Vostok. In order to estimate the 'climate error' distribution $\hat{\delta}$, we obtain values of δ at each core site from four additional experiments targeted at 2100 climate (Supplementary Information). These experiments provide an estimate of a Gaussian distribution of δ consistent with a specific change in temperature. The standard deviations are 4.6‰, 8.1‰ and 3.7‰, for Dome C, Dome F and Vostok, respectively. We assume that all errors are uncorrelated. This leads to a pessimistic error estimate for the palaeothermometer gradient because positively correlated errors tend to alter the mean value of δ rather than its gradient. Nevertheless, 95% of palaeothermometer gradient estimates, resulting from the Monte Carlo analysis, fall between 3.2‰ K^{-1} and 5.4‰ K^{-1}. These gradients lead to the range of maximum interglacial temperatures between 6.0 K and 10.0 K above present-day.

30. Uemura, R., Yoshida, N., Kurita, N., Nakawo, M. & Watanabe, O. An observation-based method for reconstructing ocean surface changes using a 340,000-year deuterium excess record from the Dome Fuji ice core, Antarctica. *Geophys. Res. Lett.* **31**, doi:10.1029/2004GL019954 (2004).

LETTERS

Reconstruction of the history of anthropogenic CO_2 concentrations in the ocean

S. Khatiwala[1], F. Primeau[2] & T. Hall[3]

The release of fossil fuel CO_2 to the atmosphere by human activity has been implicated as the predominant cause of recent global climate change[1]. The ocean plays a crucial role in mitigating the effects of this perturbation to the climate system, sequestering 20 to 35 per cent of anthropogenic CO_2 emissions[2–4]. Although much progress has been made in recent years in understanding and quantifying this sink, considerable uncertainties remain as to the distribution of anthropogenic CO_2 in the ocean, its rate of uptake over the industrial era, and the relative roles of the ocean and terrestrial biosphere in anthropogenic CO_2 sequestration. Here we address these questions by presenting an observationally based reconstruction of the spatially resolved, time-dependent history of anthropogenic carbon in the ocean over the industrial era. Our approach is based on the recognition that the transport of tracers in the ocean can be described by a Green's function, which we estimate from tracer data using a maximum entropy deconvolution technique. Our results indicate that ocean uptake of anthropogenic CO_2 has increased sharply since the 1950s, with a small decline in the rate of increase in the last few decades. We estimate the inventory and uptake rate of anthropogenic CO_2 in 2008 at 140 ± 25 Pg C and 2.3 ± 0.6 Pg C yr^{-1}, respectively. We find that the Southern Ocean is the primary conduit by which this CO_2 enters the ocean (contributing over 40 per cent of the anthropogenic CO_2 inventory in the ocean in 2008). Our results also suggest that the terrestrial biosphere was a source of CO_2 until the 1940s, subsequently turning into a sink. Taken over the entire industrial period, and accounting for uncertainties, we estimate that the terrestrial biosphere has been anywhere from neutral to a net source of CO_2, contributing up to half as much CO_2 as has been taken up by the ocean over the same period.

A key challenge for estimating anthropogenic CO_2 (C_{ant}) in the ocean is that C_{ant} is not a directly measurable quantity. Existing estimates of C_{ant} are thus based on indirect techniques, such as so-called 'back calculation' methods that attempt to separate the small anthropogenic perturbation (of the order of a few per cent) from the large background distribution of carbon by correcting the measured total dissolved inorganic carbon (DIC) concentration for changes due to biological activity and air–sea disequilibrium[5,6]. The recent availability of a high quality, global tracer data set[7] and significant improvements in methodology, notably the development of the ΔC^* approach[8], made it possible to apply these ideas, and led to one of the first observationally based global estimates of the distribution of C_{ant} in the ocean[2]. Although a major advance in our understanding of C_{ant} in the ocean, it has been suggested that this estimate suffers from a number of biases and limitations[9–11], one of them being that it provides only a snapshot for the mid-1990s.

Our method for estimating anthropogenic CO_2 builds on previous work[11,12], but extends it in several significant ways. Following previous studies, we exploit the fact that the anthropogenic perturbation can be treated as a conservative tracer[13] transported by ocean circulation from the surface mixed layer into the interior. The transport of water that carries C_{ant} into the interior ocean involves considerable dispersion and mixing of water masses of different ages and of different surface origin. Anthropogenic CO_2 at an interior location $\mathbf{x}$ and at time t is therefore related to its history in the surface mixed layer, C^s_{ant}, through a convolution equation involving a kernel, $\mathcal{G}$, which partitions each water parcel according to where and when it was last in contact with the sea surface:

$$C_{ant}(\mathbf{x}, t) = \int d\mathbf{x}' \int_{1765}^{t} dt' C^s_{ant}(\mathbf{x}', t') \mathcal{G}(\mathbf{x}, t; \mathbf{x}', t') \qquad (1)$$

The time integral in equation (1) is over the industrial era and the space integral is over the whole ocean surface. The kernel, $\mathcal{G}(x, t; x', t')$, is an intrinsic property of the ocean circulation and not of any particular tracer. As such, it can be used to propagate the surface boundary condition of any conservative tracer into the full three-dimensional concentration field. We exploit this fact by using a suite of well sampled oceanic tracers such as chlorofluorocarbons, natural ^{14}C, temperature, and salinity from the GLODAP and World Ocean Atlas databases (see Methods), to provide constraints analogous to equation (1) from which we deconvolve $\mathcal{G}$. We recognize that there is no single tracer that perfectly emulates the C_{ant} transient, necessitating the use of multiple tracers with distinct time histories to constrain $\mathcal{G}$. To regularize the under-determined deconvolution problem, we use a maximum entropy method[14] which is well suited for problems with positive kernels. We note that the C_{ant} estimated via equation (1) is relatively insensitive to errors in $\mathcal{G}$ as the latter only appears via a convolution[12].

Apart from $\mathcal{G}$, we need an estimate of C^s_{ant} in order to apply equation (1) to compute C_{ant}. We obtain this boundary condition from the known atmospheric CO_2 history by requiring that the rate of change of the inventory of C_{ant} (obtained by integrating equation (1) over the volume of the ocean) must, by mass conservation, be equal to its net flux into the ocean. The air–sea flux of C_{ant} is proportional to the change in surface disequilibrium of CO_2 (see Methods). To make further progress, we exploit the empirical result from ocean carbon cycle models that the change in disequilibrium is, to a very good approximation, proportional to the (known) anthropogenic perturbation in the atmospheric partial pressure of CO_2, p_{CO_2}. To estimate the unknown, spatially variable proportionality constant, we combine the above constraint with the requirement that our solution match observed p_{CO_2} values averaged over a discrete set of surface patches, subject to the CO_2-system equilibrium chemistry. Once the proportionality constants are known, the history of C_{ant} on each surface patch is readily obtained.

We note that our inversion method provides improvements to reduce the three main biases of most previous techniques[10], namely:

[1]Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 10964, USA. [2]Department of Earth System Science, University of California, Irvine, California 92697, USA. [3]NASA Goddard Institute for Space Studies, 2880 Broadway, New York, New York 10025, USA.

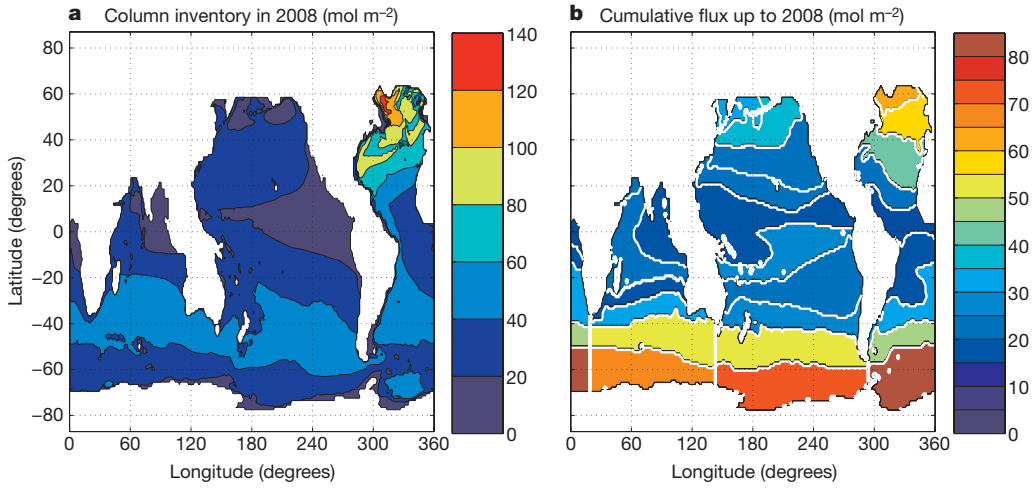


Figure 1 | Anthropogenic carbon in the ocean. a, Column inventory of C_{ant} in the ocean in 2008; the total inventory of C_{ant} in 2008 was ~140 Pg C. **b**, Cumulative C_{ant} uptake intensity up to 2008 partitioned according to surface region. The white lines delineate the 26 surface patches used in the inversion. To estimate the uncertainties quoted in the text, we repeated the analysis by randomly sampling the various parameters used in the inversion from a uniform distribution centred about the observed value of the parameter. The width of the distribution was set equal to the reported uncertainties for the tracer observations, and ±30% for the gas-transfer coefficient. Sensitivity to the prior $\mathcal{M}$ (see Methods Summary) was accounted for by similarly varying the inverse Gaussian Δ/Γ ratio between 0.5 and 1.5, and the value of Γ by ±20%. The range of C_{ant} computed using this scheme is reported as the uncertainty.

(1) the air–sea disequilibrium is allowed to evolve and is estimated by solving a nonlinear optimization problem, (2) the mixing of waters of different ages is accounted for by using multiple transient and steady tracers to constrain the age distribution, and (3) the mixing of different end-member water types is accounted for by constraining $\mathcal{G}$ using multiple steady and transient tracers. Tests using tracers simulated with a global ocean circulation model show that our method is able to successfully recover the complex spatial distribution of C_{ant} in the model with a maximum error in the column inventory of ~2 mol m^{-2} and an error in the global inventory of less than 2% (Supplementary Information).

Using our new inverse method, we have reconstructed the first three-dimensional, time-varying, history of anthropogenic carbon in the ocean from AD 1765 to AD 2008. Figure 1a shows the column inventory of C_{ant} in 2008. The total inventory in that year was ~140 ± 25 Pg C. This estimate excludes the Arctic Ocean and marginal seas not covered by the GLODAP database. Using a recent estimate based on CFC-11 for the former[15], and an area scaling approach for the latter[2], would increase our estimate of the global inventory by ~11 Pg C. We have also partitioned this inventory according to where at the surface the anthropogenic CO_2 penetrated the ocean. As is evident from Fig. 1b, the high latitude oceans, driven by intermediate and deep water formation, constitute the most intense sinks of C_{ant}. In particular, the Southern Ocean is by far the largest conduit by which anthropogenic CO_2 enters the ocean: roughly 40% of the C_{ant} residing in the ocean in 2008 entered the ocean south of 40° S.

It is useful to compare our result for 1994 to previous estimates that are available for that year. The inventory estimated using the so-called ΔC* method[2] for 1994 is 106 ± 21 Pg C. The 1994 inventory estimated using the transit-time distribution (TTD) method[11] is 107 Pg C, with a range of 94–121 Pg C. Both estimates are consistent with our 1994 estimate of 114 ± 22 Pg C. However, the previous TTD-based estimate includes a 20% downward correction for the fact that air–sea disequilibrium was incorrectly treated as being constant. Such a correction, based on model simulations, is not necessary for our estimate because our inverse method explicitly accounts for changing air–sea disequilibrium. The spatial distribution we obtain is quite different from that obtained with the ΔC* method, particularly in the Southern Ocean. Relative to the ΔC*-based estimate, our estimate of C_{ant}, like the TTD-based estimate, is generally lower in the upper ocean and higher in the deep ocean[11]. Two key aspects of the ΔC* method are the use of a single tracer age to characterize transport and the assumption of constant disequilibrium. These give rise to competing biases[10] that largely cancel out, leading to the close, but fortuitous, agreement between our estimate of the total inventory and that derived using the ΔC* method.

Figure 2 shows the uptake history over the industrial era (AD 1765 to AD 2008) computed from the time-varying inventory. (The corresponding space- and time-varying change in surface disequilibrium of CO_2 driving this uptake is also estimated by our inversion method.) There has been a sharp increase in ocean uptake since the 1950s in response to a higher growth rate of atmospheric CO_2, although the rate of increase has decreased somewhat in the last few decades. Our estimated uptake rate for the 1990s, 2.0 ± 0.6 Pg C yr^{-1}, agrees well with the IPCC consensus estimate based on independent methods[4] (Fig. 2 and Table 1).

Our work has important implications for the terrestrial carbon budget, here computed as a residual between the main fossil fuel source and the ocean and atmosphere sinks. (Including the highly

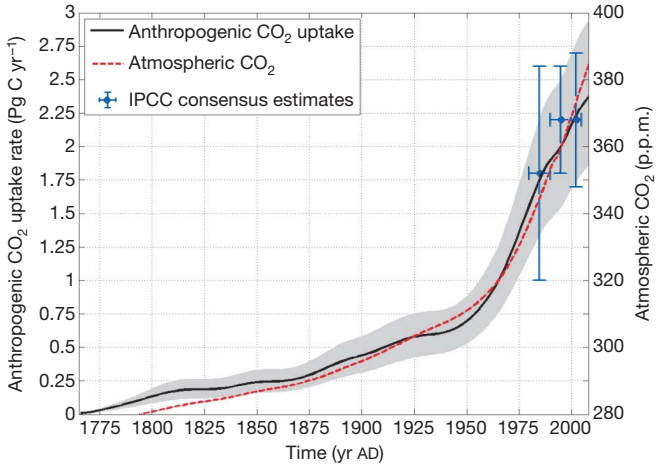


Figure 2 | Anthropogenic carbon uptake rate from 1765 to 2008 (black solid line). The shaded area represents the error envelope (see Fig. 1 legend). Also shown are the decadal average uptake rates adopted by the IPCC fourth-assessment report (AR4)[4] (blue circles; vertical error bars are ±1 s.d. and horizontal error bars span the averaging period of years) and the atmospheric CO_2 mixing ratio[29] used for the inversion (red dashed line).

Table 1 | Decadal mean ocean and land uptake rates of C_{ant}

Period	Ocean (Pg C yr^{-1})	Land (Pg C yr^{-1})	Ocean − AR4 (Pg C yr^{-1})	Land − AR4 (Pg C yr^{-1})
1980s	1.8 (1.3–2.3)	0.3 (−0.3 to 0.8)	1.8 (1.0–2.6)	0.3 (−0.6 to1.2)
1990s	2.0 (1.4–2.6)	1.1 (0.5–1.8)	2.2 (1.8–2.6)	1.0 (0.4–1.6)
2000s	2.3 (1.7–2.9)	1.1 (0.4–1.8)	2.2 (1.8–2.6)†	1.3 (0.7–1.9)†

Columns 1 and 2 show uptake rates from this work; columns 3 and 4 show corresponding values adopted by the IPCC AR4. For the current decade, the average is over 2000–06.
† IPCC ocean and land uptake values derived from a numerical ocean model simulation[17,19].

uncertain source due to changes in land use[16] provides a different perspective; Supplementary Fig. 3.) There remains considerable uncertainty regarding the relative partitioning of anthropogenic emissions between the ocean and land biosphere[3]. Our time-evolving estimate of the ocean uptake provides a more precise and detailed view of the land sink. Table 1 compares our estimates for the uptake rate of C_{ant} by the land biosphere with other estimates[4,17]. There is generally good agreement between the two, although the latter only go back to the 1980s, whereas our approach covers the entire industrial period. Figure 3 shows the evolution of the various sources and sinks of anthropogenic CO_2 between AD 1765 and AD 2005. Our results indicate that the terrestrial biosphere was a source of C_{ant} until the 1940s, roughly in line with previous model-based estimates[3,18], after which it turned into a sink of anthropogenic CO_2. Taken over the entire industrial period, and accounting for uncertainties, we estimate that the terrestrial biosphere has been anywhere from neutral to a net source of CO_2, contributing up to half as much C_{ant} as has been taken up by the ocean over the same period.

One potential source of error we have neglected is variability in ocean circulation, especially long-term trends. Such variability, however, remains poorly constrained, and its impact on anthropogenic CO_2 uptake is still debated. For example, ocean models show a slight decline in the rate-of-growth of C_{ant} uptake by the Southern Ocean over the past few decades due to an increase in the strength of the meridional overturning circulation (MOC) in response to strengthened westerly winds[19,20]. However, a recent study[21] finds no observational evidence for such a change in the MOC, suggesting that the simulated changes in the MOC and uptake may be an artefact of the eddy-parametrization used in coarse-resolution ocean models[21,22]. Nevertheless, it is useful to place these modelled changes within the context of our inverse calculations, which assume a cyclo-stationary circulation. The simulated decrease[19], of ~0.08 Pg C yr^{-1} per decade between 1981 and 2004, would imply a reduction in C_{ant} uptake of roughly 2.1 Pg C over that period (equivalent to a 1 p.p.m. increase in atmospheric CO_2 concentration). This value should be compared with a total ocean uptake of 46 ± 6 Pg C over the same period, as estimated by our inverse method (17.5 ± 2.5 Pg C for the Southern Ocean)—that is, a 5% reduction in the global ocean sink of anthropogenic CO_2, but still well within the uncertainty of our purely data-based estimate.

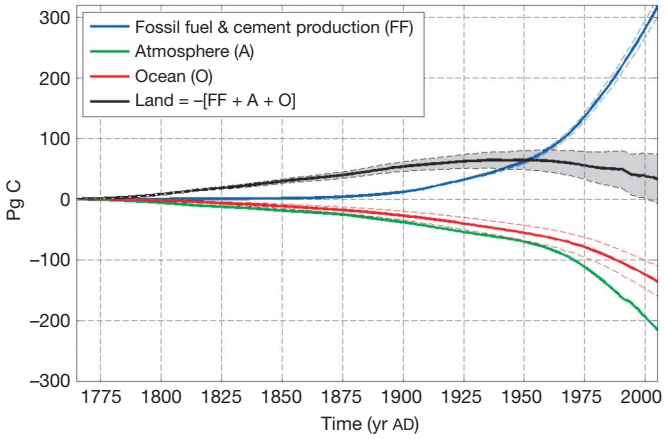


Figure 3 | Evolution of anthropogenic CO_2 sources and sinks between 1765 and 2005. Fossil fuel burning[30] (including a small contribution from cement production) is the only source considered here, and is shown as positive values. Sinks, shown as negative values, include the atmosphere, ocean, and land biosphere. Error envelope (as in Fig. 1 legend), indicated by broken lines and the shaded area, includes a 5% uncertainty in fossil fuel emissions[17].

To conclude, we have presented an observationally based estimate of the time-evolving distribution of anthropogenic CO_2 in the ocean over the industrial era. Unlike other recent inverse calculations[23], we do not rely on an ocean model for transport information. Errors in transport simulated by such models[23–25] are a large source of uncertainty in estimates of C_{ant} based on those inverse calculations. Instead, we constrain the transport by using tracer observations. Our results can thus be used to assess and improve global ocean biogeochemical models, initialize high-resolution simulations, and provide boundary conditions for atmospheric inversions of anthropogenic sources and sinks.

METHODS SUMMARY

The combined GLODAP/WOA05 databases (see Methods) provide six constraints, one for each tracer, of the form: $C(\mathbf{x}, t) = \int d\mathbf{x}' \int_{-\infty}^{t} dt' e^{-\lambda(t-t')} C(\mathbf{x}', t') \mathcal{G}(\mathbf{x}, t; \mathbf{x}', t')$, where C is the observed tracer concentration, and λ its radioactive decay rate (non-zero only for ^{14}C). To reduce the indeterminacy, we assume that the ocean circulation is stationary except for a cyclo-stationary seasonal cycle and we discretize the sea surface position variable, $\mathbf{x}'$, into a discrete set of 26 surface patches (Fig. 1b). With these assumptions, the kernel for each interior position, $\mathbf{x}$, can be written as a discrete function, $\mathcal{G}(i, k, m)$ in which i is the surface-patch index, and k and m are respectively the number of years since, and the month when, the water was last in the surface mixed layer. To further regularize the deconvolution, we used a maximum entropy approach in which we maximize an entropy functional, $J[\mathcal{G}] = -\sum_{i,k,m} \mathcal{G}(i, k, m) \log \frac{\mathcal{G}(i,k,m)}{\mathcal{M}(i,k,m)}$, subject to the discretized tracer constraints. $\mathcal{M}$ is a prior estimate of $\mathcal{G}$, which we take to be an analytical solution to the one-dimensional advection–diffusion equation known as the inverse Gaussian[26]. Simulations in ocean general circulation models[27] show that the inverse Gaussian form captures well the general characteristics of $\mathcal{G}$. The inverse Gaussian is characterized by two parameters, a mean age Γ and width Δ. Consistent with tracer observations[26,28], we set $\Delta = \Gamma$, which leaves us with one free parameter. To specify Γ, we make it a function of depth, increasing linearly from 10 yr in the surface layer to 2,000 yr in the deepest layer. The resulting variational problem is solved using the method of Lagrange multipliers, and yields a solution of the form (shown for clarity using the continuous time variables):

$$\mathcal{G}(\mathbf{x}, t; i) = \mathcal{M}(\mathbf{x}, t; i) e^{-\sum_j \alpha_j(\mathbf{x}) C_j^s(t_j - t; i) e^{-\lambda_j t}}$$

where $\alpha_j(\mathbf{x})$ is the Lagrange multiplier required to enforce the jth observational constraint. The above yields a system of nonlinear algebraic equations for the Lagrange multipliers, which we solve using standard methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 May; accepted 15 September 2009.

1. Solomon, S., *et al.* (eds) *Climate Change 2007 — The Physical Science Basis* (Cambridge Univ. Press, 2007).
2. Sabine, C. L. *et al.* The ocean sink for anthropogenic CO_2. *Science* **305,** 367–371 (2004).
3. Houghton, R. A. Balancing the global carbon budget. *Annu. Rev. Earth Planet. Sci.* **35,** 313–347 (2007).
4. Denman, K. L. *et al.* in *Climate Change 2007 — The Physical Science Basis* (eds Solomon, S. *et al.*) 499–587 (Cambridge Univ. Press, 2007).
5. Brewer, P. G. Direct measurements of the oceanic CO_2 increase. *Geophys. Res. Lett.* **5,** 997–1000 (1978).
6. Chen, C.-T. & Millero, F. J. Gradual increase of oceanic CO_2. *Nature* **277,** 205–206 (1979).
7. Key, R. M. *et al.* A global ocean carbon climatology: results from GLODAP. *Glob. Biogeochem. Cycles* **18,** doi:10.1029/2004GB002247 (2004).
8. Gruber, N., Sarmiento, J. L. & Stocker, T. F. An improved method for detecting anthropogenic CO_2 in the oceans. *Glob. Biogeochem. Cycles* **10,** 809–837 (1996).
9. Wallace, D. W. R. Ocean measurements and models of carbon sources and sinks. *Glob. Biogeochem. Cycles* **15,** 3–10, doi:10.1029/2000GB001354 (2001).
10. Matsumoto, K. & Gruber, N. How accurate is the estimation of anthropogenic carbon in the ocean? An evaluation of the ΔC* method. *Glob. Biogeochem. Cycles* **19,** doi:10.1029/2004GB002397 (2005).

11. Waugh, D. W., Hall, T. M., McNeil, B. I., Key, R. M. & Matear, R. J. Anthropogenic CO_2 in the oceans estimated using transit-time distributions. *Tellus B* **58,** 376–390 (2006).
12. Hall, T. M., Haine, T. W. N. & Waugh, D. W. Inferring the concentration of anthropogenic carbon in the ocean from tracers. *Glob. Biogeochem. Cycles* **16,** doi:10.1029/2001GB001835 (2002).
13. Sarmiento, J. L., Orr, J. C. & Siegenthaler, U. A perturbation simulation of CO_2 uptake in an ocean general circulation model. *J. Geophys. Res.* **97,** 3621–3645 (1992).
14. Tarantola, A. *Inverse Problem Theory and Methods for Model Parameter Estimation* (Society for Industrial and Applied Mathematics, 2005).
15. Tanhua, T. *et al.* Ventilation of the Arctic Ocean: mean ages and inventories of anthropogenic CO_2 and CFC-11. *J. Geophys. Res.* **114,** doi:10.1029/2008JC004868 (2009).
16. Houghton, R. A. *Carbon Flux to the Atmosphere from Land-Use Changes 1850-2005* (Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, 2008); ⟨http://cdiac.ornl.gov/trends/landuse/houghton/1850–2005.txt⟩.
17. Canadell, J. G. *et al.* Contributions to accelerating atmospheric CO_2 growth from economic activity, carbon intensity and efficiency of natural sinks. *Proc. Natl Acad. Sci. USA* **104,** 18866–18870 (2007).
18. Joos, F., Meyer, R., Bruno, M. & Leuenberger, M. The variability in the carbon sinks as reconstructed for the last 1000 years. *Geophys. Res. Lett.* **26,** 1437–1440 (1999).
19. Le Quéré, C. *et al.* Saturation of the Southern Ocean CO_2 sink due to recent climate change. *Science* **316,** doi:10.1126/science.1136188 (2007).
20. Lovenduski, N. S., Gruber, N., Doney, S. C. & Lima, I. D. Enhanced CO_2 outgassing in the Southern Ocean from a positive phase of the Southern Annular Mode. *Glob. Biogeochem. Cycles* **21,** doi:10.1029/2006GB002900 (2007).
21. Böning, C. W., Dispert, A., Visbeck, M., Rintoul, S. & Schwarzkopf, F. U. Response of the Antarctic Circumpolar Current to recent climate change. *Nature Geosci.* **1,** 864–869 (2008).
22. Zickfeld, K., Fyfe, J., Saenko, O. A., Eby, M. & Weaver, A. J. Response of the global carbon cycle to human-induced changes in the Southern Hemisphere winds. *Geophys. Res. Lett.* **34,** doi:10.1029/2006GL028797 (2007).
23. Mikaloff Fletcher, S. E. *et al.* Inverse estimates of anthropogenic CO_2 uptake, transport, and storage by the ocean. *Glob. Biogeochem. Cycles* **20,** doi:10.1029/2005GB002530 (2006).
24. Matsumoto, K. *et al.* Evaluation of ocean carbon cycle models with data-based metrics. *Geophys. Res. Lett.* **31,** doi:10.1029/2003GL018970 (2004).
25. Cao, L. *et al.* The role of ocean transport in the uptake of anthropogenic CO_2. *Biogeosciences* **6,** 375–390 (2009).
26. Waugh, D. W., Haine, T. W. & Hall, T. M. Transport times and anthropogenic carbon in the subpolar North Atlantic Ocean. *Deep Sea Res. I* **51,** 1475–1491 (2004).
27. Khatiwala, S., Visbeck, M. & Schlosser, P. Age tracers in an ocean GCM. *Deep Sea Res. I* **48,** 1423–1441 (2001).
28. Hall, T. M., Waugh, D. W., Haine, T. W. N., Robbins, P. E. & Khatiwala, S. Estimates of anthropogenic carbon in the Indian Ocean with allowance for mixing and time-varying air-sea CO_2 disequilibrium. *Glob. Biogeochem. Cycles* **18,** doi:10.1029/2003GB002120 (2004).
29. *Atmospheric Trace Gases: Carbon Dioxide* (Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, 2009); ⟨http://cdiac.ornl.gov/trends/co2⟩.
30. Boden, T. A., Marland, G. & Andres, R. J. *Global, Regional, and National Fossil-fuel CO_2 Emissions* doi:10.3334/CDIAC/00001 (Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, 2009); ⟨http://cdiac.ornl.gov/trends/emis/tre_glob.html⟩.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by US NSF grants OCE 06-23366 (to S.K. and T.H.) and OCE 07-26871 (to F.P.).

Author Contributions All authors contributed extensively to the work presented in this paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.K. (spk@ldeo.columbia.edu).

METHODS

Observational data used for deconvolution of $\mathcal{G}$. The tracers we use for the deconvolution include gridded fields of CFC-11, CFC-12, and natural ^{14}C from the GLODAP database[7], and temperature, salinity, oxygen and phosphate from the World Ocean Atlas 2005[31–34] (WOA05). While ^{14}C is not a conservative tracer, its radioactive decay rate is known and can be accounted for in the convolution. Oxygen and phosphate are also non-conservative because of the remineralization of organic matter that consumes oxygen and releases phosphate. However, the two tracers can be combined into a conservative tracer[35] $PO_4^*(\equiv PO_4 + O_2/175)$. In general, temperature, salinity and PO_4^* provide information about the mixing of different end-member water types, while the CFCs and ^{14}C provide surface-to-interior transit-time information.

To construct the surface boundary conditions for CFC-11 and CFC-12, we scaled the known atmospheric history of those tracers[36] to the measured surface concentration, hence approximately accounting for undersaturation of these gases in the mixed layer[37–39]. For the other tracers, we use a monthly mean climatology. Gas-transfer coefficients[40] were averaged over each surface patch.

We note that our inversion methodology does not require observations of carbon in the ocean interior, but does utilize surface carbon measurements (see below).

Computation of surface boundary condition for C_{ant}. To compute the boundary condition for C_{ant}, we require that the instantaneous rate of change of inventory of C_{ant} must, by mass conservation, be equal to its net flux into the ocean:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\text{volume}}\mathrm{d}\mathbf{x}\int_{\text{surface}}\mathrm{d}\mathbf{x}'\int_{1765}^{t}\mathrm{d}t' C^{s}_{ant}(\mathbf{x}',t')\mathcal{G}(\mathbf{x},t;\mathbf{x}',t') = \int F(\mathbf{x}',t)\mathrm{d}\mathbf{x}' \quad (2)$$

The air-sea flux of anthropogenic CO_2 is in turn given by

$$F(\mathbf{x}',t') = -k(\mathbf{x}')[\delta p_{CO_2}(\mathbf{x}',t') - \delta p_{CO_2}{}^{atm}(t')] \equiv -k(\mathbf{x}')\delta\Delta p_{CO_2}(\mathbf{x}',t') \quad (3)$$

where k is a gas-transfer coefficient, Δ represents the air-sea difference, and δ represents the anthropogenic perturbation. This equation shows that the flux is proportional to the change in surface disequilibrium of CO_2. We further exploit the empirical result from ocean carbon cycle models (Supplementary Information) that the change in disequilibrium is, to a very good approximation, proportional to the (known) anthropogenic perturbation in atmospheric p_{CO_2}:

$$\delta\Delta p_{CO_2}(\mathbf{x}',t') \approx \varepsilon(\mathbf{x}')\delta p_{CO_2}{}^{atm}(t') \quad (4)$$

where ε is the (unknown) proportionality constant. This allows us to recast the RHS of equation (2) in terms of ε. The LHS can similarly be recast by relating the dissolved inorganic carbon (DIC) concentration in surface waters to the partial pressure of CO_2 via the equilibrium chemistry for the CO_2 system in sea water. Denoting this equilibrium as $DIC = f(p_{CO_2})$, we have:

$$C^{s}_{ant}(\mathbf{x}',t') \approx f(p_{CO_2}(\mathbf{x}',1765) + (1+\varepsilon(\mathbf{x}'))\delta p_{CO_2}{}^{atm}(t')) - f(p_{CO_2}(\mathbf{x}',1765)) \quad (5)$$

The unknown pre-industrial surface p_{CO_2} appearing in the above equation is given by:

$$p_{CO_2}(\mathbf{x}',1765) \approx p_{CO_2}(\mathbf{x}',t_{obs}) - (1+\varepsilon(\mathbf{x}'))\delta p_{CO_2}{}^{atm}(t_{obs})$$

where $p_{CO_2}(\mathbf{x}',t_{obs})$ is the measured value of p_{CO_2} at time t_{obs}. We take these from a recently compiled database of global surface p_{CO_2} observations[41]. In this manner, the constraint equation (2) can be written entirely in terms of a single set of unknowns, the $\varepsilon(\mathbf{x}')$. In practice, since equation (2) must hold at every instant, we discretize in time with annual resolution and average in space over a discrete set of surface patches to obtain a set of nonlinear equations for the ε_i that we solve using standard nonlinear least squares.

31. Locarnini, R. A. *et al. World Ocean Atlas 2005* Vol. 1, *Temperature* (NOAA Atlas NESDIS 61, US Government Printing Office, 2006).
32. Antonov, J. I., Locarnini, R. A., Boyer, T. P., Mishonov, A. V. & Garcia, H. E. *World Ocean Atlas 2005* Vol. 2, *Salinity* (NOAA Atlas NESDIS 62, US Government Printing Office, 2006).
33. Garcia, H. E., Locarnini, R. A., Boyer, T. P. & Antonov, J. I. *World Ocean Atlas 2005* Vol. 3, *Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation* (NOAA Atlas NESDIS 63, US Government Printing Office, 2006).
34. Garcia, H. E., Locarnini, R. A., Boyer, T. P. & Antonov, J. I. *World Ocean Atlas 2005* Vol. 4, *Nutrients (Phosphate, Nitrate, Silicate)* (NOAA Atlas NESDIS 64, US Government Printing Office, 2006).
35. Broecker, W. S. *et al.* How much deep water is formed in the Southern Ocean? *J. Geophys. Res.* **103,** 15833–15844 (1998).
36. Walker, S. J., Weiss, R. F. & Salameh, P. K. Reconstructed histories of the annual mean atmospheric mole fractions for the halocarbons CFC11, CFC-12, CFC113 and carbon tetrachloride. *J. Geophys. Res.* **105,** 14285–14296 (2000).
37. Smethie, W. M. & Fine, R. A. Rates of North Atlantic Deep Water formation calculated from chlorofluorocarbon inventories. *Deep Sea Res. I* **48,** 189–215 (2001).
38. Rhein, M. *et al.* Labrador Sea Water: pathways, CFC inventory, and formation rates. *J. Phys. Oceanogr.* **32,** 648–665 (2002).
39. Lo Monaco, C., Goyet, C., Metzl, N., Poisson, A. & Touratier, F. Distribution and inventory of anthropogenic CO_2 in the Southern Ocean: comparison of three data-based methods. *J. Geophys. Res.* **110,** doi:10.1029/2004JC002571 (2005).
40. Sweeney, C. *et al.* Constraining global air-sea gas exchange for CO_2 with recent bomb ^{14}C measurements. *Glob. Biogeochem. Cycles* **21,** doi:10.1029/2006GB002784 (2007).
41. Takahashi, T. *et al.* Climatological mean and decadal change in surface ocean pCO_2, and net sea-air CO_2 flux over the global oceans. *Deep Sea Res. II* **56,** doi:10.1016/j.dsr2.2008.12.009 (2009).

Vol 462|19 November 2009|doi:10.1038/nature08496

LETTERS

Mutation load and rapid adaptation favour outcrossing over self-fertilization

Levi T. Morran[1], Michelle D. Parmenter[1] & Patrick C. Phillips[1]

The tendency of organisms to reproduce by cross-fertilization despite numerous disadvantages relative to self-fertilization is one of the oldest puzzles in evolutionary biology. For many species, the primary obstacle to the evolution of outcrossing is the cost of production of males[1], individuals that do not directly contribute offspring and thus diminish the long-term reproductive output of a lineage. Self-fertilizing ('selfing') organisms do not incur the cost of males and therefore should possess at least a twofold numerical advantage over most outcrossing organisms[2]. Two competing explanations for the widespread prevalence of outcrossing in nature despite this inherent disadvantage are the avoidance of inbreeding depression generated by selfing[3–5] and the ability of outcrossing populations to adapt more rapidly to environmental change[1,6,7]. Here we show that outcrossing is favoured in populations of *Caenorhabditis elegans* subject to experimental evolution both under conditions of increased mutation rate and during adaptation to a novel environment. In general, fitness increased with increasing rates of outcrossing. Thus, each of the standard explanations for the maintenance of outcrossing are correct, and it is likely that outcrossing is the predominant mode of reproduction in most species because it is favoured under ecological conditions that are ubiquitous in natural environments.

The vast majority of animals and plants reproduce by outcrossing, as opposed to selfing. This observation is puzzling, because theory suggests selfing enjoys several substantial fitness advantages over outcrossing[8,9]. For example, selfing results in the production of offspring that are each capable of bearing offspring, whereas many outcrossing species produce males that do not bear offspring. This halving of the number of offspring-bearing progeny an individual can produce is known as the 'two-fold cost of males' and generates a large gap between the mating systems in numerical contribution, and thus fitness, over time[1]. In addition to this inherent numerical advantage, selfing also efficiently reduces the mutation load over time by eliminating or 'purging' new harmful mutations by exposing them to natural selection via the production of homozygous offspring[3,4]. However, if mutations are too numerous or have effect sizes that allow them to slip below the selection threshold, then deleterious mutations can accumulate unchecked within selfing lineages—something that should not happen in outcrossing populations of sufficient size[5,10,11]. Further, any new adaptive mutations will tend to become trapped within different selfing lineages because the lack of outcrossing means that any mutations that arise within separate selfing individuals can not be incorporated into the same lineage or genome[12,13]. In this way, selfing mimics the problems associated with asexual reproduction, with outcrossing providing a more effective means of recombination and thereby generating the genetic variation necessary to adapt to a novel environment[7]. To critically evaluate these theoretical predictions, it is necessary both to experimentally manipulate the mating system of a given species and to recapitulate the evolutionary process under the specific conditions predicted to favour either selfing or outcrossing.

Here, we utilize experimental evolution in populations of *C. elegans* to test the benefits of outcrossing relative to selfing under conditions predicted to favour outcrossing. *C. elegans* populations are composed of males and hermaphrodites. Hermaphrodites reproduce through either self-fertilization or by outcrossing with males. Despite the potential for outcrossing with males, most *C. elegans* populations reproduce predominantly via selfing ('wild-type' outcrossing rates are generally less than 5%)[14–19]. However, by incorporating one of two mating system altering mutations (*xol-1* and *fog-2*; refs 20 and 21, respectively), we generated both obligate selfing and obligate outcrossing populations, yielding three different outcrossing levels (obligate selfing, wild type, obligate outcrossing) within the same genetic background. These mutations were independently crossed into two separate genetic backgrounds (N2 and CB4856) with known differences in wild-type outcrossing rates[19]. Exposing these populations to two different novel selection environments—(1) elevated mutation rates coupled with a migratory barrier (Supplementary Fig. 1a) and (2) a virulent bacterial pathogen (Supplementary Fig. 1b)—allowed us to directly test theories advocating either deleterious mutations or adaptation to ecological conditions as the primary selective forces contributing to the prevalence of outcrossing as a means of sexual reproduction.

Selfing populations are thought to be able to purge new deleterious mutations as long as the mutations are not too frequent and their effect sizes are large enough to be exposed to selection[3–5]. Indeed, even relatively small *C. elegans* populations have been shown to escape the most serious consequences of mutation accumulation, even when their mutation rate is increased tenfold[22]. However, outcrossing is predicted to slow the fixation of deleterious mutations with weak to moderate effect sizes. To explore these contrasting expectations, we subjected populations to the chemical mutagen ethyl methanesulphonate (EMS) every other generation at a level that increases individual mutation rate by approximately four times the natural rate. Populations exposed to the mutagen and populations maintained at natural mutation rates were reared and passaged within a novel environment (a Petri dish transected by a vermiculite barrier separating populations from their food source upon introduction to the dish) to impose strong selection and thereby facilitate the potential to purge deleterious mutations. We then tracked the subsequent evolution of 60 different populations for 50 generations under different combinations of mutation, mating system and genetic background.

Despite strong selection against deleterious mutations, obligate selfing populations fixed significantly more mutations than did the obligate outcrossing populations, as evidenced by the fact that the latter populations maintained fitness over the course of the experiment in spite of elevated mutation rates, whereas the selfing populations displayed a substantial decline in fitness (Fig. 1a; analysis of variance: $F_{1,481} = 456.15$, $P < 0.001$). The purging of deleterious mutations within selfing

[1]Center for Ecology & Evolutionary Biology, 5289 University of Oregon, Eugene, Oregon 97403-5289, USA.

populations is easily overwhelmed by slight increases in mutation rate. In contrast, while outcrossing populations are more likely to accumulate segregating deleterious mutations[11], these mutations do not lead to an overall decline in mean fitness (Fig. 1a). The value of outcrossing is particularly evident in the wild-type populations, where outcrossing rates are free to vary as dictated by selection. The wild-type populations subject to elevated mutation rates exhibit increased levels of outcrossing (Fig. 1b; $F_{1,8} = 55.7$, $P < 0.001$), indicating that increased levels of outcrossing are favoured under these conditions.

Whereas fitness loss due to selfing is offset to a large extent by the intermediate amounts of outcrossing exhibited in the wild-type populations, obligate selfing CB4856 populations lose fitness over time even when maintained at their natural mutation rate (Fig. 1a; $F_{1,481} = 17.5$, $P < 0.001$). We replicated the deterministic loss of fitness in obligate selfing CB4856 populations under long term maintenance in more permissive laboratory conditions as well (20% fitness loss over 30 generations; $F_{3,71} = 9.85$, $P < 0.001$). Indeed, obligate selfing *C. elegans* populations would in general be expected to go extinct over the course of a few hundred generations[23]. Several other studies have investigated the role that elevated mutation rates may play in maintaining males within partially selfing *C. elegans* populations, finding that increases in mutation can prolong the maintenance of males in the population, but at levels that are only slightly greater than wild type[24,25]. Therefore, even partial outcrossing is a valuable, if not always sufficient, means of managing the influx of deleterious mutations.

As predicted, outcrossing ameliorates the fixation of deleterious mutations. However, alternative theories emphasize that outcrossing should enable a stronger and more rapid adaptive response to ecological conditions than selfing[1,6,7,12,13]. Here, outcrossing (wild-type and obligate outcrossing) populations maintained at natural mutation rates exhibited a significantly greater amount of adaptation than the obligate selfing populations after 50 generations of selection, regardless of genetic background (Fig. 1a; $F_{1,481} = 51.98$, $P < 0.001$). The observed rate of adaptation in the obligate outcrossing populations (0.34% increase in fitness per generation) is particularly impressive, because this adaptation occurred in near-isogenic lines over a span of only 50 generations. Thus, the majority of the adaptive response is likely to have been due to novel mutations.

To further test the ability of outcrossing to facilitate rapid adaptation, we exposed obligate outcrossing, wild type, and obligate selfing populations within a common CB4856 background to the bacterial pathogen *Serratia marcescens*. Several strains of *S. marcescens* elicit a pathogen avoidance behaviour from *C. elegans*[26], in addition to inducing the expression of a specific set of pathogen resistance genes following ingestion[27]. *S. marcescens* 2170 is highly virulent when consumed by *C. elegans*, initially inducing an 80% mortality rate in our experimental regime (Supplementary Fig. 1b). Repeated exposures to *S. marcescens* therefore impose strong selection for either pathogen avoidance or resistance, or a combination of both responses. As a control, replicate populations were passaged on heat-killed *S. marcescens*. Before selection on *S. marcescens*, the experimental populations were mutagenized with EMS to generate standing genetic variation into the previously inbred experimental populations.

After 40 generations of exposure to *S. marcescens*, outcrossing populations adapted to the novel pathogenic conditions whereas the obligate selfing populations did not (Fig. 1c; $F_{1,80} = 245.79$, $P < 0.001$). The obligate outcrossing populations exhibited very rapid and substantial increases in fitness when exposed to *S. marcescens* (Fig. 1c; $F_{1,80} = 160.18$, $P < 0.001$). In addition, wild-type mating populations exposed to *S. marcescens* exhibited elevated outcrossing rates (Fig. 1d;

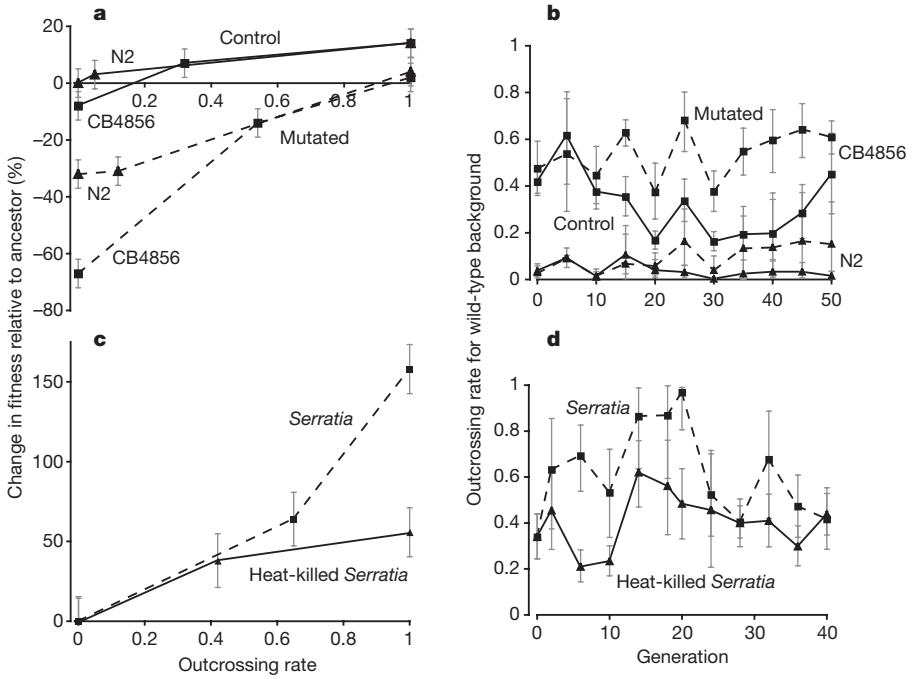


Figure 1 | Experimental test of the major theories of the evolution of outcrossing. a, Experimental populations (N2, triangles; CB4856, squares) with different outcrossing rates were exposed to a novel, challenging environment at either natural (solid lines) or elevated (4×; dashed lines) mutation rates for 50 generations. Percentage change in population mean fitness over time was assessed by comparing the competitive fitness of the ancestral population to that of the evolved population. Obligately selfing populations showed pronounced fitness decline in the face of elevated mutation rates (or even natural mutation rates in the case of CB4856). Both the rate of adaptation and resistance to mutational degradation increased with increasing levels of outcrossing. **b**, Within the wild-type outcrossing treatments, populations exposed to elevated mutation rates evolved higher outcrossing rates. **c**, Experimental populations with a CB4856 background were mutated to generate genetic variation and then exposed to either the bacterial pathogen *S. marcescens* (dashed lines) or heat-killed *S. marcescens* control (solid lines) for 40 generations, then the percentage change in mean fitness was measured for each population. The outcrossing populations exhibited both rapid and substantial adaptation to the pathogen, but the obligate selfing populations failed to adapt. **d**, Populations exposed to *S. marcescens* evolved higher outcrossing rates within the wild-type outcrossing treatment. Thus, in keeping with theory, both the influx of deleterious mutations and adaptation to a novel environment favour outcrossing over selfing. Points represent the means of 5 replicate experimental evolution populations for **a** and **b**, and 6–7 populations for **c** and **d**. Error bars, ±2 s.e.m. (errors calculated on arcsine-square-root transformed data for **b** and **d**).

$F_{1,5} = 27.2$, $P = 0.003$) and significantly greater fitness (Fig. 1c; $F_{1,80} = 9.29$, $P = 0.003$) than wild-type populations maintained on heat-killed *S. marcescens*, indicating that selection favoured outcrossing over selfing. In general, outcrossing first increased and then declined over the course of the experiment (approaching its maximum value of 1.0 after 20 generations), indicating that the change in outcrossing is an evolved rather than facultative response (Fig. 1d). Stronger selection imposed by *S. marcescens* and initial standing genetic variation enabled a much stronger evolutionary response (3.8% increase in fitness per generation) (Fig. 1c) than that observed in the first experiment (Fig. 1a). Overall, then, outcrossing enables more rapid adaptation to changing ecological conditions than does selfing.

The prevalence of outcrossing is something of an evolutionary puzzle, given the inherent advantages of self-fertilization. This work provides the first (to our knowledge) experimental tests of the selective pressures favouring the evolution and maintenance of outcrossing. We have demonstrated that outcrossing impedes the fixation of deleterious mutations and facilitates rapid adaptation relative to selfing, such that outcrossing is at least conditionally favoured by selection. Similar results have been observed in accelerated rates of evolutionary change in sexual versus asexual populations[28,29]. Although we cannot directly address the question of the origin of selfing and outcrossing in our experiments, overall levels of outcrossing increased in our wild-type treatments in which selfed and outcrossed offspring were competing within the same population (Fig. 1b, d). These results support the idea that obligate selfing may often be an evolutionary dead-end, in which species that evolve obligate selfing are ultimately doomed to extinction owing to an inability to respond to changing environmental conditions[6].

The fact that obligate outcrossing yielded a much larger response than natural outcrossing rates is something of a surprise, because it is thought that moderate amounts of outcrossing are sufficient to escape the problems associated with obligate selfing[11]. One additional feature of this system that has not been previously considered, however, is that an increase in the frequency of males within a population also increases the opportunity for sexual selection, which has been shown to reduce the overall genetic load within a population[30]. Males therefore play multiple roles within these populations, both for enhancing genetic exchange across generations and increasing the efficacy of natural selection within generations. Mutation, changing environmental conditions, and pathogens are nearly ubiquitous selective pressures for many organisms, which probably explains outcrossing's relative prevalence in nature.

METHODS SUMMARY

We conducted two large-scale experimental evolution studies. First, we exposed obligate outcrossing, wild-type mating, and obligate selfing populations with approximately 500 individuals apiece to 0.5 mM EMS every other generation for 50 generations. These mutated populations, in addition to replicate populations maintained at natural mutation rates, were passaged each generation in a selective novel environment (Supplementary Fig. 1a). Second, we exposed obligate outcrossing, wild-type mating, and obligate selfing populations composed of approximately 500 individuals to *S. marcescens* (Supplementary Fig. 1b) for 40 generations while exposing replicate populations to heat-killed *S. marcescens* as a control. These populations were exposed to 10 mM EMS for four generations before selection as a means of inducing genetic variation. We used a competitive fitness assay to measure the change in fitness for each experimental population relative to its ancestor before selection. The competitive fitness assays were conducted within the context of the selective environment and the assay was conducted simultaneously on the experimental population and the previously frozen ancestral population. Fitness was determined by mixing each population (experimental and ancestral) with a GFP-marked tester strain at a 50:50 ratio. After passaging the worms in the relevant selective environment, the GFP ratio of the offspring was calculated and used to estimate fitness.

Received 13 August 2009; accepted 11 September 2009.
Published online 21 October 2009.

1. Maynard Smith, J. *The Evolution of Sex* (Cambridge Univ. Press, 1978).
2. Lively, C. M. & Lloyd, D. G. The cost of biparental sex under individual selection. *Am. Nat.* **135,** 489–500 (1990).
3. Charlesworth, D. & Charlesworth, B. Inbreeding depression and its evolutionary consequences. *Annu. Rev. Ecol. Syst.* **18,** 237–268 (1987).
4. Lande, R. & Schemske, D. W. The evolution of self-fertilization and inbreeding depression in plants. 1. Genetic models. *Evolution* **39,** 24–40 (1985).
5. Heller, J. & Maynard Smith, J. Does Muller's Ratchet work with selfing? *Genet. Res.* **8,** 269–294 (1979).
6. Stebbins, G. L. Self-fertilization and population variation in higher plants. *Am. Nat.* **91,** 337–354 (1957).
7. Crow, J. F. An advantage of sexual reproduction in a rapidly changing environment. *J. Hered.* **83,** 169–173 (1992).
8. Fisher, R. A. Average excess and average effect of a gene substitution. *Ann. Eugen.* **11,** 53–63 (1941).
9. Williams, G. C. *Sex and Evolution* (Princeton Univ. Press, 1975).
10. Kondrashov, A. S. Deleterious mutations as an evolutionary factor. I. The advantage of recombination. *Genet. Res.* **44,** 199–217 (1984).
11. Schultz, S. T. & Lynch, M. Mutation and extinction: the role of variable mutational effects, synergistic epistasis, beneficial mutations, and degree of outcrossing. *Evolution* **51,** 1363–1371 (1997).
12. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78,** 737–756 (1974).
13. Barton, N. H. Linkage and the limits to natural selection. *Genetics* **140,** 821–841 (1995).
14. Chasnov, J. R. & Chow, K. L. Why are there males in the hermaphroditic species *Caenorhabditis elegans? Genetics* **160,** 983–994 (2002).
15. Stewart, A. D. & Phillips, P. C. Selection and maintenance of androdioecy in *Caenorhabditis elegans. Genetics* **160,** 975–982 (2002).
16. Sivasundar, A. & Hey, J. Population genetics of *Caenorhabditis elegans*: the paradox of low polymorphism in a widespread species. *Genetics* **163,** 147–157 (2003).
17. Barriere, A. & Felix, M. A. High local genetic diversity and low outcrossing rate in *Caenorhabditis elegans* natural populations. *Curr. Biol.* **15,** 1176–1184 (2005).
18. Haber, M. *et al.* Evolutionary history of *Caenorhabditis elegans* inferred from microsatellites: evidence for spatial and temporal genetic differentiation and the occurrence of outbreeding. *Mol. Biol. Evol.* **22,** 160–173 (2005).
19. Teotónio, H., Manoel, D. & Phillips, P. C. Genetic variation for outcrossing among *Caenorhabditis elegans* isolates. *Evolution* **60,** 1300–1305 (2006).
20. Miller, L. M., Plenefisch, J. D., Casson, L. P. & Meyer, B. J. *xol-1* — a gene that controls the male modes of both sex determination and X-chromosome dosage compensation in *C. elegans. Cell* **55,** 167–183 (1988).
21. Schedl, T. & Kimble, J. *fog-2*, a germ-line-specific sex determination gene required for hermaphrodite spermatogenesis in *Caenorhabditis elegans. Genetics* **119,** 43–61 (1988).
22. Estes, S., Phillips, P. C., Denver, D. R., Thomas, W. K. & Lynch, M. Mutation accumulation in populations of varying size: the distribution of mutational effects for fitness correlates in *Caenorhabditis elegans. Genetics* **166,** 1269–1279 (2004).
23. Loewe, L. & Cutter, A. D. On the potential for extinction by Muller's ratchet in *Caenorhabditis elegans. BMC Evol. Biol.* **8,** 125 (2008).
24. Cutter, A. D. Mutation and the experimental evolution of outcrossing in *Caenorhabditis elegans. J. Evol. Biol.* **18,** 27–34 (2005).
25. Manoel, D., Carvalho, S., Phillips, P. C. & Teotónio, H. Selection against males in *Caenorhabditis elegans* under two mutational treatments. *Proc. R. Soc. Lond. B* **274,** 417–424 (2007).
26. Pradel, E. *et al.* Detection and avoidance of a natural product from the pathogenic bacterium *Serratia marcescens* by *Caenorhabditis elegans. Proc. Natl Acad. Sci. USA* **104,** 2295–2300 (2007).
27. Mallo, G. V. *et al.* Inducible antibacterial defense system in *C. elegans. Curr. Biol.* **12,** 1209–1214 (2002).
28. Colegrave, N. Sex releases the speed limit on evolution. *Nature* **420,** 664–666 (2002).
29. Goddard, M. R., Charles, H., Godfray, J. & Burt, A. Sex increases the efficacy of natural selection in experimental yeast populations. *Nature* **434,** 636–640 (2005).
30. Whitlock, M. C. & Agrawal, A. F. Purging the genome with sexual selection: reducing mutation load through selection on males. *Evolution* **63,** 569–582 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Scholz, A. Ohdera and J. Chiem for logistical help, S. Katz for providing the *S. marcescens* 2170 strain, and J. Thornton for use of laboratory space and equipment. We also thank B. Cresko, C. Lively, J. Thornton and the members of the Phillips and Cresko laboratories for comments and discussion pertaining to this work. Funding was provided by NSF grants DEB-0236180, DEB-0710386 and DEB-0641066, and an NIH Genetics Fellowship awarded to L.T.M. Some nematode strains used in this work were provided by the *Caenorhabditis* Genetics Center, which is funded by the NIH National Center for Research Resources (NCRR).

Author Contributions L.T.M. and P.C.P. designed the experiments. L.T.M. and M.D.P. performed the experiments. L.T.M. and P.C.P. analysed the data. L.T.M. and P.C.P. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.C.P. (pphil@uoregon.edu).

LETTERS

Frequency of gamma oscillations routes flow of information in the hippocampus

Laura Lee Colgin[1], Tobias Denninger[1]†, Marianne Fyhn[1]†, Torkel Hafting[1]†, Tora Bonnevie[1], Ole Jensen[2], May-Britt Moser[1] & Edvard I. Moser[1]

Gamma oscillations are thought to transiently link distributed cell assemblies that are processing related information[1,2], a function that is probably important for network processes such as perception[1–3], attentional selection[4] and memory[5,6]. This 'binding' mechanism requires that spatially distributed cells fire together with millisecond range precision[7,8]; however, it is not clear how such coordinated timing is achieved given that the frequency of gamma oscillations varies substantially across space and time, from ~25 to almost 150 Hz[1,9–13]. Here we show that gamma oscillations in the CA1 area of the hippocampus split into distinct fast and slow frequency components that differentially couple CA1 to inputs from the medial entorhinal cortex, an area that provides information about the animal's current position[14–17], and CA3, a hippocampal subfield essential for storage of such information[14,18,19]. Fast gamma oscillations in CA1 were synchronized with fast gamma in medial entorhinal cortex, and slow gamma oscillations in CA1 were coherent with slow gamma in CA3. Significant proportions of cells in medial entorhinal cortex and CA3 were phase-locked to fast and slow CA1 gamma waves, respectively. The two types of gamma occurred at different phases of the CA1 theta rhythm and mostly on different theta cycles. These results point to routeing of information as a possible function of gamma frequency variations in the brain and provide a mechanism for temporal segregation of potentially interfering information from different sources.

We investigated the function of gamma frequency variations in the hippocampus, a medial temporal lobe structure that plays a critical role in memory[20]. Hippocampal gamma oscillations are thought to arise from two sources, one in the entorhinal cortex (EC)[9,21] and another intrinsic to the hippocampus[9,10]. The estimated current sources during hippocampal gamma oscillations closely match the currents that result from stimulation of the perforant path projection from EC to the hippocampus[9], indicating that hippocampal gamma may be entrained by direct inputs from EC. Entorhinal gamma has been reported to be relatively fast (~90 Hz)[22], and high-frequency gamma (~80 Hz) has been reported also in the hippocampus[9]. However, in animals with EC lesions, a slower gamma rhythm (~40 Hz) becomes more apparent in the hippocampus. The pattern of current dipoles for this slower oscillation matches the current profile associated with activation of the Schaffer collateral/commissural pathway from CA3 to CA1[9]. Collectively, these observations indicate that hippocampal gamma oscillations have multiple origins and raise the possibility that variations in gamma frequency in CA1 reflect alternating synchronization with slow gamma in CA3 and fast gamma in EC (Supplementary Fig. 1). To test this idea, we sampled neural activity simultaneously from CA1 and either CA3 or layer III of medial entorhinal cortex (MEC) in freely moving rats.

Gamma oscillations in CA1 had two distinct frequency components, a slow gamma range (~25–50 Hz) and a fast gamma range (~65–140 Hz) (Fig. 1A, B and Supplementary Figs 2–4). For 22 CA1 stratum pyramidale recordings from 16 rats, we quantified the coupling between theta phase and power in the gamma frequency range using cross-frequency analyses, which provide a sensitive measure of the coupling between theta phase and gamma power[12]. We found that theta phase was coupled to power in two separate bands of gamma in nearly all experiments (Fig. 1C and Supplementary Figs 5a, b and 6). Individual slow and fast gamma episodes were extracted by selecting periods when power within the trial-specific slow and fast gamma frequency bands exceeded 2 s.d. above the mean slow and fast gamma power, respectively. The analyses showed that slow and fast gamma episodes were associated with different portions of the underlying theta cycle (Fig. 1D and E, panel a), although both appeared within the half-cycle of theta when gamma power is maximal[9,12]. Slow gamma amplitude peaked during the early descending part of the theta wave ($257 \pm 66^\circ$, mean phase $\pm$ angular deviation; 0° was defined as the trough of the theta cycle in stratum pyramidale). Fast gamma was maximal near the trough ($329 \pm 63^\circ$), closely aligned with the theta phase when the CA1 pyramidal cells were most active ($0 \pm 71^\circ$; Fig. 1D, bottom panel). The phase difference between slow and fast gamma maxima was significant ($F(2,20) = 7.6$, $P < 0.01$; two-tailed Hotelling test for paired samples of angles) whereas the phase difference between fast gamma and maximal number of spikes was not ($F(1,189) = 3.8$, $P > 0.1$, Watson–Williams test for two samples of angles).

Within individual theta cycles, gamma power was typically concentrated in either the slow or fast gamma range but not both (Fig. 1E). Significant negative correlations ($P < 0.001$) between slow and fast gamma power on individual theta cycles were observed in all 22 CA1 recordings (16 rats). Gamma oscillations of either type were detected on over half ($54 \pm 17\%$) of all CA1 theta cycles (means $\pm$ s.d.). On these theta cycles, the fast variant of gamma was detected on a higher proportion of theta cycles than the slow variant (fast gamma: $77 \pm 11\%$; slow gamma: $45 \pm 9\%$; Fig. 1E, panels b, c), with a small percentage exhibiting power that exceeded threshold in both slow and fast gamma frequency ranges ($21 \pm 5\%$; Fig. 1E, panel d). Gamma episodes of the same subtype tended to cluster together (Supplementary Fig. 7).

Slow and fast gamma episodes in CA1 were compared to electroencephalograms (EEG) from corresponding windows of time in simultaneous recordings from layer III of MEC (Fig. 2a) or from CA3 (Fig. 2b). During periods of slow gamma in CA1, little to no synchronous gamma activity was seen in MEC. In contrast, during periods of fast gamma in CA1, substantial fast gamma co-occurred in MEC.

[1]Kavli Institute for Systems Neuroscience and Centre for the Biology of Memory, MTFS, Olav Kyrres gate 9, Norwegian University of Science and Technology, NO-7489 Trondheim, Norway. [2]Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, P.O. Box 9101, Nijmegen NL-6500 HB, The Netherlands. †Present addresses: Massachusetts Institute of Technology, The Picower Institute for Learning and Memory, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA (T.D.); University of California San Francisco, Department of Physiology, 513 Parnassus Avenue, San Francisco, California 94143, USA (M.F. and T.H.).

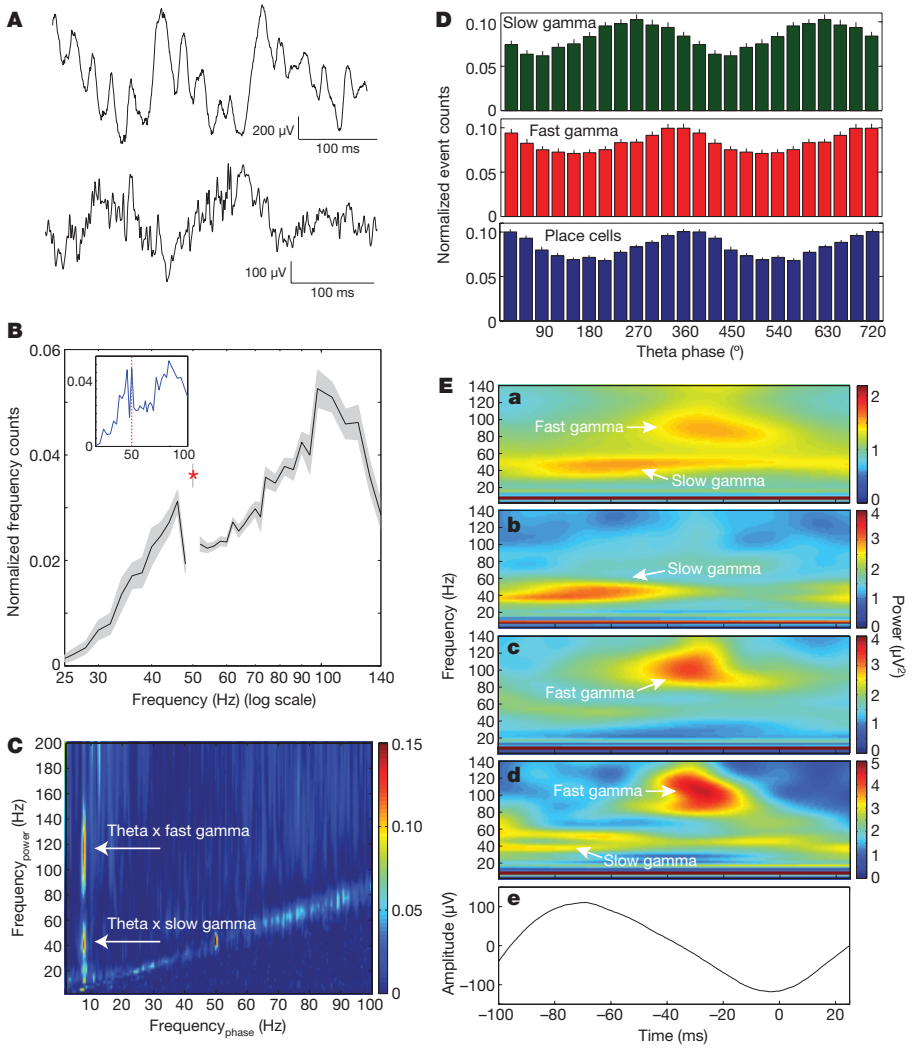


Figure 1 | Two bands of gamma oscillations in CA1. A, Example raw recordings during periods of high power slow gamma (top) and fast gamma (bottom) oscillations. Gamma rhythms are seen riding on slower and larger theta waves. **B**, Distribution of instantaneous gamma frequencies (mean ± s.e.m.) shows two peaks, separated at ~55 Hz. Because no 50 Hz filters were applied, the decline of the slow gamma peak was followed by a sharp 50 Hz noise peak (red asterisk). The separation between the slow gamma and the 50 Hz peak is also shown for an individual example (inset). **C**, Example cross-frequency coherence plot showing that two separate bands of gamma power (*y* axis) were modulated by theta phase (*x* axis). Coupling strength (C) is colour-coded (dark blue, no coupling; red, maximal coupling). **D**, Theta phase distributions (means ± s.e.m.) are shown for slow gamma maxima (top), fast gamma maxima (middle), and place cell spike times (bottom) for 22 CA1 recordings from 16 rats. 0° was defined as the trough of the theta cycle. **E**, Time frequency representations of power for a representative recording, averaged across (**a**) all theta cycles, (**b**) theta cycles with slow gamma, (**c**) theta cycles with fast gamma and (**d**) the minority of theta cycles exhibiting both slow and fast gamma. Power is colour-coded (a prewhitening filter was applied for this illustration to equalize power at gamma and theta frequencies). Frequency is plotted on the *y* axis, and time is depicted on the *x* axis with $t = 0$ corresponding to the theta trough. (**e**) The averaged unfiltered theta cycles.

Across the group of six rats with simultaneous MEC–CA1 recordings, the power of fast gamma in MEC during fast gamma in CA1 was significantly higher, relative to simultaneous CA1 gamma power measures, than the power of slow gamma in MEC during slow gamma in CA1 (Fig. 2a, bottom right panel; two-tailed paired *t*-test, $t(5) = 2.5$, $P < 0.05$; Supplementary Fig. 8). In the group of 11 rats with paired recordings from CA3 and CA1, slow gamma in CA1 was accompanied by prominent slow gamma activity in CA3 (Fig. 2b). Some fast gamma was also observed in CA3 during fast gamma in CA1, but its power relative to gamma power in simultaneous CA1 recordings was significantly smaller than the power of slow gamma in CA3 during slow gamma in CA1 (Fig. 2b, bottom right panel; two-tailed paired *t*-test, $t(10) = 7.4$, $P < 0.0001$). Together, these results demonstrate that CA1 synchronizes with CA3 during slow gamma and with MEC during fast gamma.

We further quantified oscillatory coupling between CA1 and layer III of MEC and between CA1 and CA3 by estimating fast and slow gamma coherence between these regions (Fig. 3). Paired recordings from MEC and CA1 showed substantial coherence in the fast gamma band, but little coherence for slow gamma frequencies (Fig. 3a). Fast gamma coherence was greater than slow gamma coherence in all MEC–CA1 cases (Fig. 3b; two-tailed paired *t*-test, six recordings from six rats, $t(5) = 4.3$, $P < 0.007$). Paired recordings from CA3 and CA1 showed a different pattern. Gamma coherence between CA3 and CA1 was concentrated in the slow gamma range (Fig. 3c; Supplementary Fig. 9). Slow gamma coherence was greater than fast gamma coherence in all paired CA3–CA1 recordings (Fig. 3d; two-tailed paired *t*-test, 17 recordings from 11 rats, $t(16) = 6.5$, $P < 0.0001$). A significant interaction effect (recording sites × type of gamma) was obtained for slow and fast gamma coherence values from the MEC–CA1 and CA3–CA1 paired recordings (two-way repeated measures ANOVA, $F(1,21) = 38.5$, $P < 0.0001$). In one animal, we recorded simultaneously from all three regions and found that CA1 displayed significant gamma phase synchrony with MEC only in the fast gamma range and with CA3 only in the slow gamma range (Supplementary Fig. 10). These observations support the conclusion

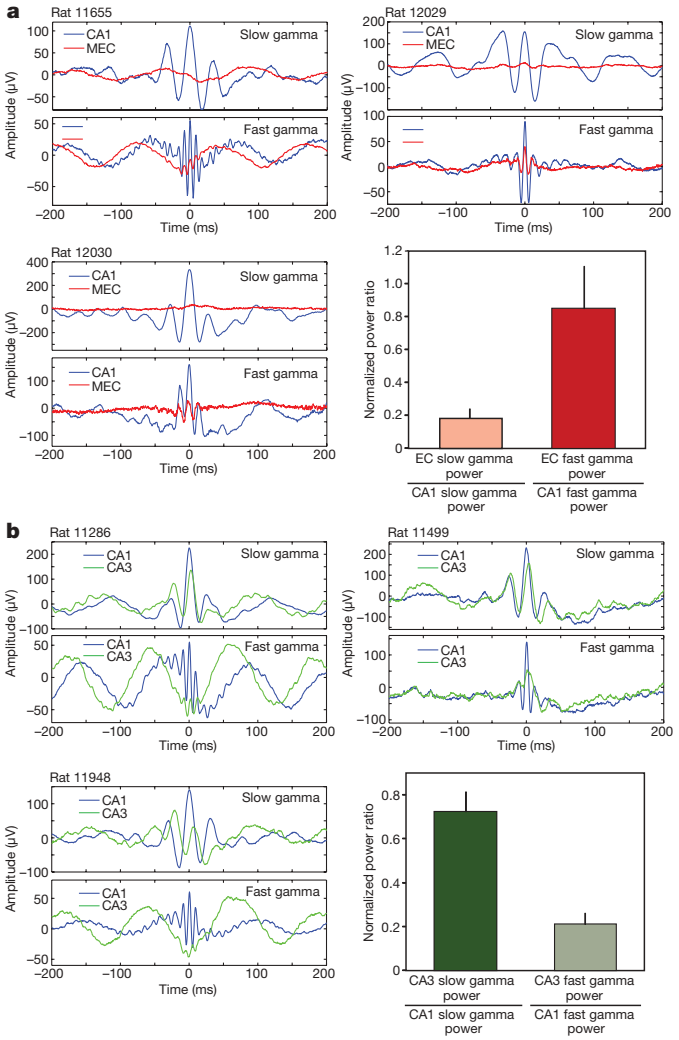


Figure 2 | Differential coupling with MEC and CA3 during fast and slow gamma oscillations in CA1. a, Representative examples from three rats show averaged recordings from layer III of MEC and CA1 during slow and fast gamma episodes detected in CA1. $t = 0$ corresponds to the peak amplitudes of the gamma episodes in CA1. The bottom right panel presents mean (±s.e.m.) EC:CA1 slow and fast gamma power ratios for all six rats with simultaneous CA1–MEC recordings. **b**, Averaged recordings of slow and fast gamma episodes detected in CA1 and corresponding windows of time from CA3 are shown for 3 example rats. The bottom right panel shows mean (±s.e.m.) CA3:CA1 slow and fast gamma power ratios for the group of 11 rats with simultaneous CA3–CA1 recordings.

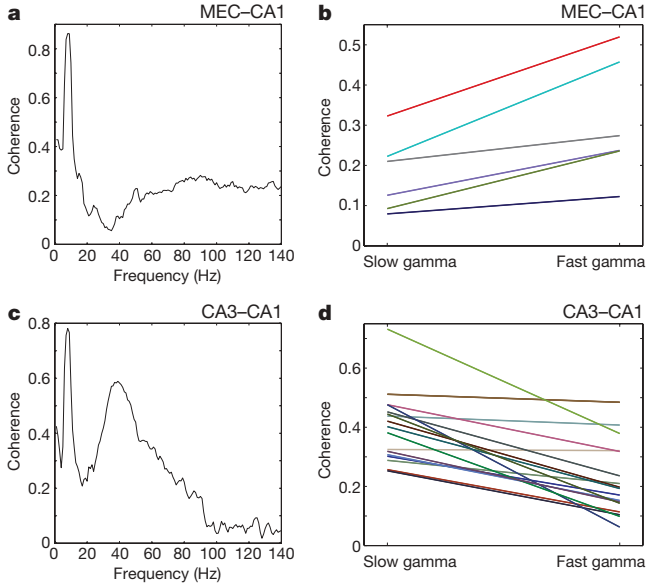


Figure 3 | Interregional synchrony during slow and fast gamma oscillations. a, Coherence plot for a representative pair of MEC–CA1 recordings. The largest peak is observed at theta frequency, but marked coherence is also apparent in the fast gamma range (~60–140 Hz). Little coherence is seen in the slow gamma range (~25–55 Hz). The relatively sharp peak at 50 Hz reflects electrical noise. **b**, Fast gamma coherence was greater than slow gamma coherence for all MEC–CA1 recording pairs. **c**, Coherence plot for a representative pair of CA3–CA1 recordings. Note that coherence in the theta frequency band is followed by a pronounced peak in the slow gamma range (~40 Hz). Little gamma coherence is observed in the fast gamma range. **d**, Slow gamma coherence was higher than fast gamma coherence for all CA3–CA1 recording pairs.

that MEC–CA1 communication is enhanced during fast gamma in CA1 and that CA3–CA1 transmission is heightened during slow gamma in CA1. Significant differences in spike time correlations between regions further support this conclusion (Supplementary Fig. 11).

To determine if fast gamma synchrony between MEC and CA1 was associated with changes in cell firing properties in MEC, we assessed the firing patterns of 29 putative principal cells and three putative interneurons from layer III in MEC during slow and fast gamma in CA1 (Supplementary Fig. 12; Supplementary Tables 1 and 2). Forty-four per cent of the entorhinal cells were significantly locked to the phase of fast gamma in CA1 (Fig. 4a; Supplementary Table 1). No MEC cells were locked to CA1 slow gamma phase (binomial test for fast against slow gamma, $P < 0.0001$; Supplementary Table 3).

To investigate whether slow gamma synchrony between CA3 and CA1 was reflected in the firing patterns of CA3 cells, we recorded 107 CA3 place cells and seven putative CA3 interneurons and analysed their spiking with respect to slow and fast gamma phase in CA1 (Fig. 4b, c; Supplementary Tables 1 and 2; Supplementary Fig. 12). The proportion of CA3 cells that were phase-locked to CA1 slow gamma was significantly higher than the proportion of cells locked to CA1 fast gamma (53% compared with 32%, binomial test, $P < 0.02$). The phase-locking of some CA3 cells to fast gamma in CA1 (Fig. 4b; Supplementary Fig. 12) may result from fast gamma coherence between CA3 and layer II of MEC (Supplementary Fig. 13) together with synchronized gamma across layers II and III in EC[22].

In the final analysis, we investigated the phase of firing of CA1 cells with respect to slow and fast gamma oscillations in CA1 (Fig. 4d, 4e). We recorded a total of 169 CA1 place cells and 17 putative CA1 interneurons (Supplementary Tables 1 and 2; Supplementary Fig. 12). Place-specific firing was observed during both slow and fast gamma epochs but the firing fields were larger during slow gamma (Supplementary Fig. 14 and 15). A significantly higher proportion of CA1 place cells were phase-locked to CA1 fast gamma than to CA1 slow gamma (38% compared with 25%, binomial test, $P < 0.03$). Only 6% of CA1 place cells showed significant phase-locking to both slow and fast gamma (Supplementary Table 1). The fast gamma-modulated place cells tended to spike near the gamma trough and the slow gamma-modulated place cells preferred to fire closer to the gamma peak (Fig. 4d, 4e; Supplementary Fig. 12). These observations indicate that slow and fast gamma-modulated place cells correspond to separate populations of CA1 neurons firing at different gamma phases and carrying different temporal codes[23].

The main finding of this study is that CA1 exhibits two distinct frequency bands of gamma oscillations that selectively synchronize CA1 with different sources of afferent input (Supplementary Fig. 1). Slow gamma synchronizes CA1 with CA3, and fast gamma synchronizes CA1 with MEC. Considering that gamma synchronization facilitates interactions between brain regions[24], the results indicate that fast gamma enhances transmission from MEC to CA1 and that slow gamma promotes signalling from CA3 to CA1.

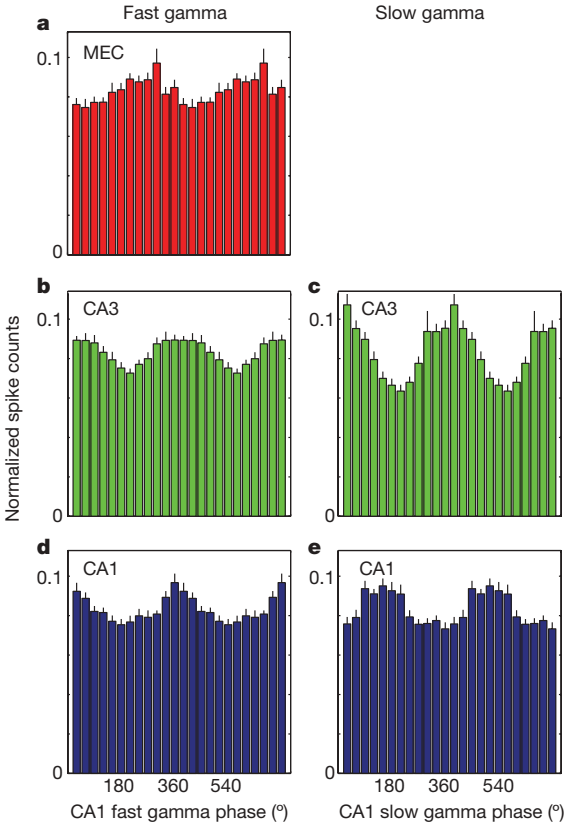


Figure 4 | Cell firing patterns during fast and slow gamma oscillations. Distributions of spike times across phases of fast and slow gamma in CA1 (means ± s.e.m.) for cells with significant phase-locking. **a**–**e**, The distribution of **a**, fast gamma firing phases for MEC principal cells; **b**, fast gamma firing phases for CA3 place cells; **c**, slow gamma firing phases for CA3 place cells; **d**, fast gamma firing phases for CA1 place cells; and **e**, slow gamma firing phases for CA1 place cells. No MEC cells were phase-locked to slow CA1 gamma.

The data support the notion that direct input from layer III of MEC is important for activating CA1 place cells, particularly in the centre of their firing field. A significantly higher proportion of place cells in CA1 was driven by fast gamma from MEC than by slow gamma from CA3. Additionally, fast gamma power in CA1 was maximal near the theta trough, the theta phase when CA1 place cells are most likely to fire. These observations fit well with previous studies showing that place-selective firing in CA1 depends on direct inputs from layer III of MEC[14,17] and indicate that transmission of spatial information between these regions is facilitated by fast gamma oscillations. Considering that the power of fast gamma oscillations in the prefrontal and parietal cortices has recently been found to be modulated by CA1 theta phase[13], the fast gamma-mediated coupling between CA1 and MEC may extend further to other cortical regions.

A significantly lower percentage of CA1 place cells was phase-locked to slow gamma than to fast gamma. Slow gamma occurs primarily at a theta phase when CA1 place cells fire with relatively low probability and is probably driven by feedforward inhibition from CA3 that transiently suppresses CA1 firing[10]. Perhaps more easily able to overcome inhibition during slow gamma are cell ensembles with synapses that had previously undergone long-term potentiation, a process that lastingly strengthens the responses of neurons to inputs and is believed to underlie memory storage[25]. During periods of slow gamma, CA1 place cells were activated across larger spatial regions, indicating that the cells fired at earlier stages of trajectories through the firing fields, possibly as a consequence of long-term potentiation of CA3–CA1 synapses[26,27]. Together, these findings raise the possibility that slow gamma conveys information from memory stores in the CA3 or CA3–CA1 network[18,19]. In line with such an idea, coherence between gamma activity in CA3 and CA1 is increased during retrieval of spatial information in a hippocampus-dependent memory task[6].

The results are consistent with previous studies reporting that inputs from EC and CA3 arrive in CA1 at different phases of the theta cycle[28]. Long-term potentiation in CA1 is most easily induced at a particular phase of theta[29], corresponding to the phase when EC input is maximal[28]. This indicates that the theta phase when EC inputs preferentially arrive may coincide with the time when memory encoding occurs optimally and raises the possibility that the EC-coupled CA1 fast gamma observed in the present study serves to facilitate memory encoding[30]. Retrieval of information is thought to occur at a different theta phase than memory encoding, during which time CA3 input to CA1 is maximal and incoming signals from EC are suppressed[28]. This idea fits well with the above-hypothesized memory retrieval function for slow gamma. Separation of afferent inputs to CA1 on different phases of theta is probably important for avoiding re-encoding of previously stored memories and also for reliably distinguishing perceptions of ongoing experiences from internally evoked memories. The present results raise the possibility that slow and fast gamma play an important role in this separation of inputs by filtering out improperly timed signals from one afferent while facilitating transfer of coherent activity from another. Considering that broadband gamma oscillations occur in other areas[1–4,11,24], separation of gamma oscillations into discrete frequency channels may be used throughout the brain to enhance interregional communication.

METHODS SUMMARY

Field potentials and neuronal activity were recorded in 29 male Long Evans rats implanted with a 'hyperdrive' ($n = 12$ for the main study; $n = 9$ for additional analyses) or a pair of 'microdrives' ($n = 4$ for the main study; $n = 4$ for additional analyses). Hyperdrives contained 14 independently movable tetrodes assembled in one or two bundles. Microdrives consisted of bundles of 4 tetrodes each. In all animals of the main study, tetrodes were implanted above the dorsal hippocampus and/or MEC (Supplementary Fig. 16), such that activity could subsequently be recorded simultaneously from CA1 and either CA3 (10 rats) or MEC (5 rats), or from all three areas (1 rat). When EEG characteristics (sharp wave polarity, theta modulation) or behavioural firing patterns (place fields, grid fields, head direction tuning) indicated that the target region had been reached, spike-triggered activity and EEG were sampled in blocks of 10–30 min while the animal explored a familiar environment. For further details, see the Supplementary Methods section.

For each recording session in CA1, slow and fast gamma frequency bands were defined on the basis of cross-frequency coherence analyses of the local EEG. Slow and fast gamma oscillatory periods with power of 2 s.d. above the mean were extracted from the CA1 recordings and corresponding windows of time were collected from simultaneous CA3 and/or MEC recordings. Coherence was computed for paired recordings from MEC–CA1 and CA3–CA1. EEG results were not affected by different sampling frequencies or high-cut filter settings (Supplementary Figs 17 and 18, see also Supplementary Methods). For spikes that occurred during slow and/or fast gamma, slow and/or fast gamma phase at the time of firing was estimated. Circular statistics were applied to test if cells were significantly phase-locked to slow or fast gamma ($P < 0.05$). Additional details about analyses are provided online in the Methods and Supplementary Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 July; accepted 9 October 2009.

1. Gray, C. M., Konig, P., Engel, A. K. & Singer, W. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338,** 334–337 (1989).
2. Fries, P., Nikolic, D. & Singer, W. The gamma cycle. *Trends Neurosci.* **30,** 309–316 (2007).
3. Freeman, W. J. Spatial properties of an EEG event in the olfactory bulb and cortex. *Electroencephalogr. Clin. Neurophysiol.* **44,** 586–605 (1978).
4. Fries, P., Reynolds, J. H., Rorie, A. E. & Desimone, R. Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* **291,** 1560–1563 (2001).
5. Lisman, J. E. & Idiart, M. A. Storage of 7 +/− 2 short-term memories in oscillatory subcycles. *Science* **267,** 1512–1515 (1995).

6. Montgomery, S. M. & Buzsaki, G. Gamma oscillations dynamically couple hippocampal CA3 and CA1 regions during memory task performance. *Proc. Natl Acad. Sci. USA* **104,** 14495–14500 (2007).
7. von der Malsburg, C. in *Models of Neural Networks II* (eds van Hemmen, J. L. & Hordemann, G. J.) 95–119 (Springer, 1994).
8. Engel, A. K., Fries, P. & Singer, W. Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Rev. Neurosci.* **2,** 704–716 (2001).
9. Bragin, A. *et al.* Gamma (40–100 Hz) oscillation in the hippocampus of the behaving rat. *J. Neurosci.* **15,** 47–60 (1995).
10. Csicsvari, J., Jamieson, B., Wise, K. D. & Buzsaki, G. Mechanisms of gamma oscillations in the hippocampus of the behaving rat. *Neuron* **37,** 311–322 (2003).
11. Kay, L. M. Two species of gamma oscillations in the olfactory bulb: dependence on behavioral state and synaptic interactions. *J. Integr. Neurosci.* **2,** 31–44 (2003).
12. Canolty, R. T. *et al.* High gamma power is phase-locked to theta oscillations in human neocortex. *Science* **313,** 1626–1628 (2006).
13. Sirota, A. *et al.* Entrainment of neocortical neurons and gamma oscillations by the hippocampal theta rhythm. *Neuron* **60,** 683–697 (2008).
14. Brun, V. H. *et al.* Place cells and place recognition maintained by direct entorhinal-hippocampal circuitry. *Science* **296,** 2243–2246 (2002).
15. Fyhn, M., Molden, S., Witter, M. P., Moser, E. I. & Moser, M. B. Spatial representation in the entorhinal cortex. *Science* **305,** 1258–1264 (2004).
16. Hafting, T., Fyhn, M., Molden, S., Moser, M. B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436,** 801–806 (2005).
17. Brun, V. H. *et al.* Impaired spatial representation in CA1 after lesion of direct input from entorhinal cortex. *Neuron* **57,** 290–302 (2008).
18. Sutherland, R. J., Whishaw, I. Q. & Kolb, B. A behavioural analysis of spatial localization following electrolytic, kainite- or colchicines-induced damage to the hippocampal formation in the rat. *Behav. Brain Res.* **7,** 133–153 (1983).
19. Steffenach, H. A., Sloviter, R. S., Moser, E. I. & Moser, M. B. Impaired retention of spatial memory after transaction of longitudinally oriented axons of hippocampal CA3 pyramidal cells. *Proc. Natl Acad. Sci. USA* **99,** 3194–3198 (2002).
20. Squire, L. R., Stark, C. E. & Clark, R. E. The medial temporal lobe. *Annu. Rev. Neurosci.* **27,** 279–306 (2004).
21. Charpak, S., Pare, D. & Llinas, R. The entorhinal cortex entrains fast CA1 hippocampal oscillations in the anaesthetized guinea-pig: role of the monosynaptic component of the perforant path. *Eur. J. Neurosci.* **7,** 1548–1557 (1995).
22. Chrobak, J. J. & Buzsaki, G. Gamma oscillations in the entorhinal cortex of the freely behaving rat. *J. Neurosci.* **18,** 388–398 (1998).
23. Senior, T. J., Huxter, J. R., Allen, K., O'Neill, J. & Csicsvari, J. Gamma oscillatory firing reveals distinct populations of pyramidal cells in the CA1 region of the hippocampus. *J. Neurosci.* **28,** 2274–2286 (2008).
24. Womelsdorf, T. *et al.* Modulation of neuronal interactions through neuronal synchronization. *Science* **316,** 1609–1612 (2007).
25. Martin, S. J., Grimwood, P. D. & Morris, R. G. Synaptic plasticity and memory: an evaluation of the hypothesis. *Annu. Rev. Neurosci.* **23,** 649–711 (2000).
26. Blum, K. I. & Abbott, L. F. A model of spatial map formation in the hippocampus of the rat. *Neural Comput.* **8,** 85–93 (1996).
27. Mehta, M. R., Barnes, C. A. & McNaughton, B. L. Experience-dependent, asymmetric expansion of hippocampal place fields. *Proc. Natl Acad. Sci. USA* **94,** 8918–8921 (1997).
28. Hasselmo, M. E., Bodelon, C. & Wyble, B. P. A proposed function for hippocampal theta rhythm: separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Comput.* **14,** 793–817 (2002).
29. Huerta, P. T. & Lisman, J. E. Bidirectional synaptic plasticity induced by a single burst during cholinergic theta oscillation in CA1 *in vitro*. *Neuron* **15,** 1053–1063 (1995).
30. Jutras, M. J., Fries, P. & Buffalo, E. A. Gamma-band synchronization in the macaque hippocampus and memory formation. *J. Neurosci.* **29,** 12521–12531 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. M. Amundsgaard, K. Haugen, K. Jenssen, E. Sjulstad, R. Skjerpeng and H. Waade for technical assistance, M. P. Witter for assistance with recording site localization, E. J. Henriksen and K. Jezek for donating rats for supplementary analyses, C. A. Barnes for helpful comments on the manuscript and G. Buzsáki and a number of other colleagues for helpful discussions. This work was supported by the Kavli Foundation and a Centre of Excellence grant from the Norwegian Research Council.

Author Contributions L.L.C., O.J., M.-B.M. and E.I.M. planned experiments and analyses, L.L.C., T.D., M.F., T.H. and T.B. collected data, L.L.C., T.D. and O.J. wrote analysis programs, L.L.C. and T.D. analysed data, and L.L.C. and E.I.M. wrote the paper, in collaboration with M.-B.M. All authors discussed the results.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to L.L.C. (laura.colgin@ntnu.no) or E.I.M. (edvard.moser@ntnu.no).

METHODS

Cross-frequency analysis (Fig. 1C and Supplementary Fig. 5). The time-varying power in a particular frequency band of the EEG (from 2–200 Hz, in 2 Hz wide frequency bins) was calculated using a Morlet's wavelets technique (see Supplementary Methods). Coherence was estimated between the original signal and the time-varying power of the signal for each frequency of interest using the magnitude squared coherence function (Hanning window with 50% overlap; 16384 and 8192 FFT points for Axona and Neuralynx data, respectively) from the Signal Processing Toolbox in MATLAB (MathWorks). This provided a measure for determining if power changes at particular frequencies (for example, gamma) were correlated with the phase at other frequencies (for example, theta).

Selection of slow and fast gamma bandwidths. Results of cross-frequency coherence analyses were used to determine, for each CA1 recording, the gamma frequency bands in which the cross-frequency coherence values exceeded mean coherence between theta phase and gamma power. The bands were rounded to multiples of 5 Hz. The lower limit for gamma was set to 25 Hz in order to separate the oscillations from theta harmonic, which could extend beyond 20 Hz in some cases. The upper limit was set to 140 Hz to avoid contamination from single spikes because spike waveforms were associated with power across a wide band of frequencies beginning around 150–200 Hz (Supplementary Fig. 3; note that power magnitudes were greatly amplified for illustration purposes via application of a strong prewhitening filter). To rule out contributions from single spikes in the field potential, cross-frequency analyses were also performed for recordings in stratum radiatum and stratum lacunosum-moleculare, where the EEG did not contain individual spike waveforms (Supplementary Fig. 5b). In most cases, and in recordings from both the cell layer and the dendritic layers, a clear separation between theta-modulated slow and fast gamma was observed, with slow gamma extending up to ~55 Hz and fast gamma beginning at ~60 Hz or higher (Fig. 1C and Supplementary Fig. 5ab). In animals in which distinct bands for either slow or fast gamma or both could not be separated using this method (for example, number 12029 in Supplementary Fig. 5a), broad bands were used with a split point that was selected on the basis of the minimum peak frequency in the distribution of gamma frequency peaks for the entire group of recordings (that is, mode of frequency distribution minima occurred at ~55 Hz, see Fig. 1B and Supplementary Fig. 4; 25–55 Hz frequency limits for slow gamma and 60–140 Hz frequency limits for fast gamma). The slow and fast gamma frequency bands defined for each trial were used in subsequent analyses.

Detection of gamma episodes. To extract periods of slow and fast gamma activity in the EEG, we computed time-varying power (see Supplementary Materials and Methods) within the frequency bands for slow and fast gamma set for each CA1 recording. Power at each time point was averaged across the slow and fast gamma frequency ranges to obtain the time-varying estimates of slow and fast gamma power. Time points were collected when slow and fast gamma power exceeded 2 s.d. above the time-averaged power of slow and fast gamma, respectively. Time windows, 160 ms in length, were cut around the selected time points. In each 160 ms segment, the slow and fast gamma amplitude maxima were determined from the slow and fast gamma bandpass filtered versions of the recordings. Duplicated gamma periods, a common consequence of extracting overlapping time windows, were avoided by discarding identical maxima values within a given gamma subtype and further requiring that maxima of a given subtype be separated by at least 100 ms. Individual gamma windows were finally constructed from the original, non-bandpass filtered recordings as 400 ms long windows centred around the slow and fast gamma amplitude maxima.

Relationship of gamma to theta phase (Fig. 1D). EEG recordings were bandpass filtered in the theta range (6–10 Hz), and theta phases for each time point were estimated using the Hilbert transform function from the Signal Processing Toolbox in MATLAB. Theta phases at the time points associated with slow and fast gamma maxima (determined as described above) as well as theta phases for CA1 place cell spikes were collected. Theta phases for each type of event (slow gamma maxima, fast gamma maxima, CA1 place cell spikes) were sorted into 30 degree bins, allowing the phase distribution of each event to be determined. For a given recording, the distributions of slow gamma, fast gamma and spikes were normalized by dividing the bins by the total number of corresponding events (that is, total number of slow gamma episodes, fast gamma episodes, or spikes within a given recording). The normalized distributions were averaged across recordings. In this analysis, and all analyses involving oscillation phase, the oscillation trough was defined as 0°.

Gamma power ratios. Power ratios (Fig. 2a, b, bottom right panels) were computed for slow and fast gamma episodes in CA1 and the corresponding windows of time collected from MEC and/or CA3. The power spectra for individual slow and fast gamma windows in CA1 and corresponding windows of time in MEC and CA3 were estimated using Welch's method (50% overlapping windows, Hanning tapers and 512 fast Fourier transform points) from the Signal Processing Toolbox in MATLAB. Power was averaged across slow gamma frequencies for slow gamma power and fast gamma frequencies for fast gamma power. The individual power spectra were normalized by dividing power at each frequency by the sum of power across all frequencies, and for each recording in each region, the normalized slow and fast gamma power measures were averaged across episodes. Slow and fast gamma power ratios were then computed by dividing the normalized slow or fast gamma power in CA3 or MEC by the corresponding normalized slow or fast gamma power in CA1.

Coherence. For each pair of EEG recordings (CA1 and MEC or CA1 and CA3), coherence (Fig. 3) was averaged across the slow gamma frequencies to estimate slow gamma coherence and across the fast gamma frequencies to estimate fast gamma coherence (see Supplementary Materials and Methods).

LETTERS

Systems-level dynamic analyses of fate change in murine embryonic stem cells

Rong Lu[1]†, Florian Markowetz[2]*†, Richard D. Unwin[3]*, Jeffrey T. Leek[4]†, Edoardo M. Airoldi[2]†, Ben D. MacArthur[4,5], Alexander Lachmann[5], Roye Rozov[4]†, Avi Ma'ayan[5], Laurie A. Boyer[6], Olga G. Troyanskaya[2], Anthony D. Whetton[3] & Ihor R. Lemischka[1,4]

Molecular regulation of embryonic stem cell (ESC) fate involves a coordinated interaction between epigenetic[1–4], transcriptional[5–10] and translational[11,12] mechanisms. It is unclear how these different molecular regulatory mechanisms interact to regulate changes in stem cell fate. Here we present a dynamic systems-level study of cell fate change in murine ESCs following a well-defined perturbation. Global changes in histone acetylation, chromatin-bound RNA polymerase II, messenger RNA (mRNA), and nuclear protein levels were measured over 5 days after downregulation of Nanog, a key pluripotency regulator[13–15]. Our data demonstrate how a single genetic perturbation leads to progressive widespread changes in several molecular regulatory layers, and provide a dynamic view of information flow in the epigenome, transcriptome and proteome. We observe that a large proportion of changes in nuclear protein levels are not accompanied by concordant changes in the expression of corresponding mRNAs, indicating important roles for translational and post-translational regulation of ESC fate. Gene-ontology analysis across different molecular layers indicates that although chromatin reconfiguration is important for altering cell fate, it is preceded by transcription-factor-mediated regulatory events. The temporal order of gene expression alterations shows the order of the regulatory network reconfiguration and offers further insight into the gene regulatory network. Our studies extend the conventional systems biology approach to include many molecular species, regulatory layers and temporal series, and underscore the complexity of the multi-layer regulatory mechanisms responsible for changes in protein expression that determine stem cell fate.

We applied a single well-defined perturbation to murine ESCs by downregulating Nanog, a key pluripotency factor[13–15]. A lentiviral-based complementation system was introduced into mouse ESCs in which short hairpin RNA (shRNA) depletes endogenous *Nanog* mRNA, and normal levels of Nanog expression are restored in a doxycycline-dependent manner from an shRNA 'immune' version[7] (Fig. 1b). Previously, we showed that this engineered ESC clone is fully pluripotent *in vitro* and *in vivo* when maintained in the presence of doxycycline[7]. After doxycycline removal, Nanog mRNA and protein levels rapidly decline (Fig. 1c), and both pluripotency and self-renewal capacities of ESCs diminish with time. We collected data from four molecular layers. Specifically, we performed: (1) chromatin-immunoprecipitation microarray (ChIP-chip) analysis of histone H3 lysine 9 and 14 acetylation (acH3K9/14) at gene promoter regions to assess chromatin modification (designated as HIS); (2) ChIP-chip analysis of RNA polymerase II localization at 3′ exons of gene coding regions to reveal active transcription (designated as POL); (3) gene expression microarrays to quantify mRNA abundance (designated as RNA); and (4) protein mass spectrometry to measure nuclear protein abundance (designated as PRO) (Fig. 1a). Fold changes were calculated for each gene by comparing the expression levels of a molecular layer on days 1, 3 and 5 (doxycycline absent, Nanog depleted) to day 0 (doxycycline present, Nanog expressing), allowing for comparisons across the different experimental platforms (Supplementary Fig. 1). To estimate experimental noise, a significance threshold in each experiment was determined based on the experimental replicates of all measured genes at a false discovery rate (FDR) of 5% (Fig. 1d and Supplementary Fig. 2).

Although changes between different gene expression steps are generally correlated (Supplementary Fig. 3), both concordances and discordances exist on the individual gene level. The discordances show regulatory events that alter gene expression. We performed a supervised gene/protein classification to identify the key regulatory step that is most responsible for changes in protein levels, which directly determine cellular phenotype. We anchored our analysis on observed changes in protein levels and assessed the concordance of changes in the other three layers by comparing PRO to RNA, then RNA to POL, and finally POL to HIS (Fig. 2a). Proteins with significant changes were assigned to one of four categories at each time-point: category 1 proteins exhibit discordant PRO and RNA changes in expression, which is indicative of translational and posttranslational regulation; category 2 proteins exhibit concordant PRO and RNA changes in expression, but discordant RNA and POL changes in expression, which is indicative of post-transcriptional regulation; category 3 proteins exhibit concordant PRO, RNA and POL changes in expression, but discordant POL and HIS changes in expression, which is indicative of transcriptional regulation; and category 4 proteins exhibit concordant changes in expression across all four layers, which is indicative of regulation through chromatin modification. Proteins tend to stay in the same category over time (Supplementary Fig. 4). Category 1 constitutes 43–52% of all the genes with significant changes in protein levels, indicating that

[1]Department of Molecular Biology, [2]Lewis-Sigler Institute for Integrative Genomics and Department of Computer Science, Princeton University, Princeton, New Jersey 08544, USA. [3]Stem Cell and Leukaemia Proteomics Laboratory, School of Cancer and Imaging Sciences, Manchester Academic Health Science Centre, University of Manchester, Wolfson Molecular Imaging Centre, Manchester M20 4QL, UK. [4]Department of Gene and Cell Medicine and The Black Family Stem Cell Institute, [5]Department of Pharmacology and System Therapeutics and Systems Biology Center New York (SBCNY), Mount Sinai School of Medicine, New York, New York 10029, USA. [6]Massachusetts Institute of Technology, Department of Biology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. †Present addresses: Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Beckman Center B261, 279 Campus Drive, Stanford, California 94305, USA (R.L.); Cancer Research UK, Cambridge Research Institute, Cambridge CB2 0RE, UK (F.M.); Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics 615 North Wolfe Street, Baltimore, Maryland 21205, USA (J.T.L.); Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, Massachusetts 02128, USA (E.M.A.); Blavatnik School of Computer Science, Tel Aviv University, 69978 Tel Aviv, Israel (R.R.).
*These authors contributed equally to this work.

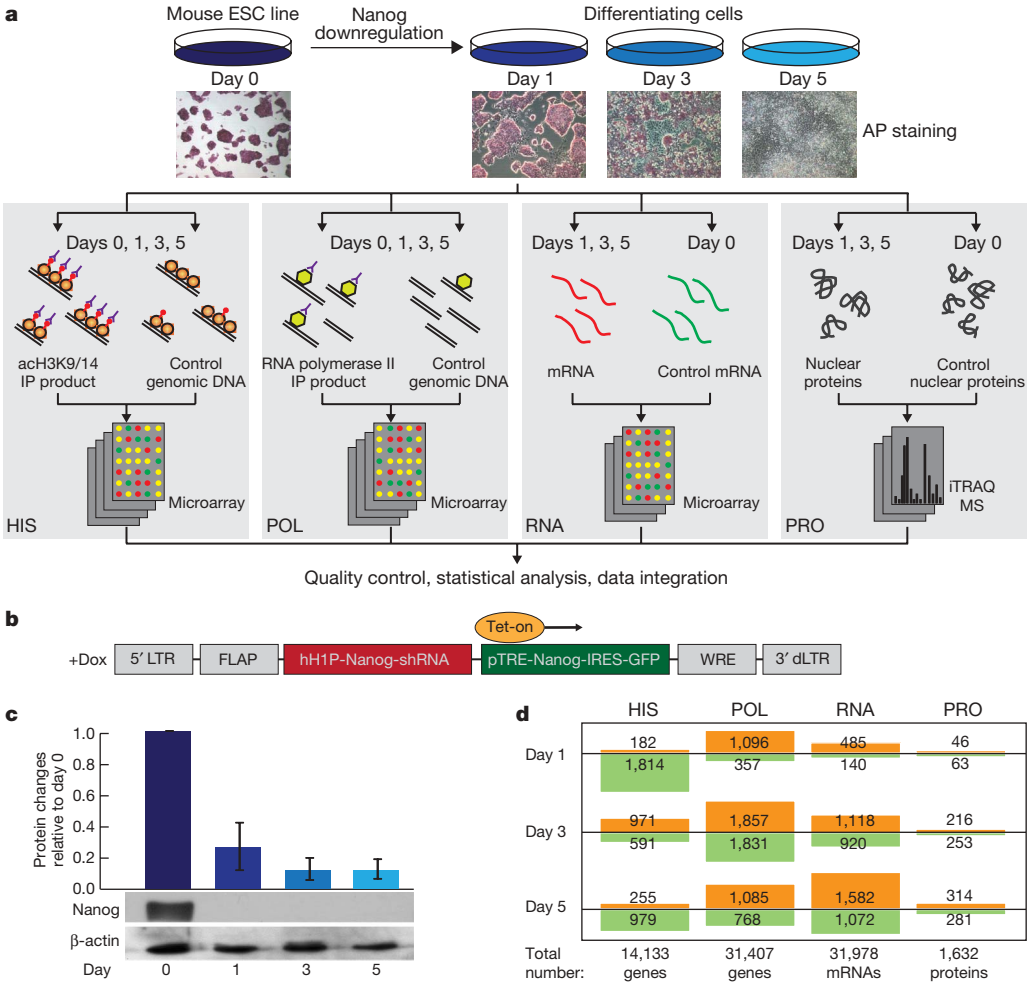


Figure 1 | Measuring changes in the epigenome, the transcriptome and the nuclear proteome after Nanog downregulation. **a**, Experimental design. AP, alkaline phosphatase; IP, immunoprecipitation; iTRAQ, isobaric tag for relative and absolute quantification; MS, mass spectrometry. **b**, The lentiviral vector construct to conditionally regulate Nanog expression levels[7]. dLTR, deleted long-terminal repeat; FLAP, nucleotide segment that improves transduction efficiency; Tet-on, tetracycline transactivator; WRE, woodchuck hepatitis virus post-transcriptional regulatory element. **c**, Efficacy of Nanog protein downregulation as measured by mass spectrometry (bar chart) and western blot (image, bottom). Error bars denote the s.d. of duplicate measurements. **d**, Summary of the numbers of genes with significant changes at different molecular layers on each day. Increased and decreased levels are shown in orange and green, respectively.

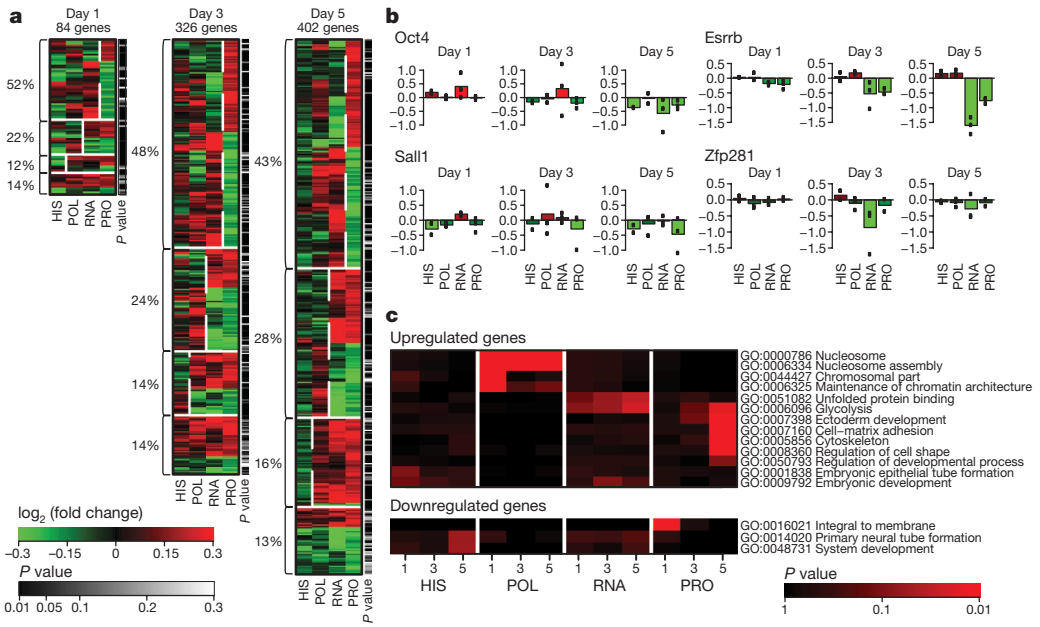


Figure 2 | Comparisons across different molecular regulatory layers. **a**, Proteins with significant changes on each day are assigned to one of four categories on the basis of concordance between expression steps (Methods). The percentages on the left are calculated according to the number of proteins in each category. The *P*-value bar on the right gives the inclusion significance level. **b**, Examples of proteins from each of the four categories. Black dots represent the exact values for each experimental replicate. **c**, Selected gene-ontology (GO) categories that are overrepresented at each gene expression step. The complete panel is shown in Supplementary Fig. 5.

translational and post-translational regulatory mechanisms have important roles in ESC fate decisions[11,12,16,17]. However, it is unclear whether this is specific to stem cells or whether it is characteristic of other biological systems.

In addition to providing a genome-wide perspective of ESC fate change, our concordance analysis also provides useful information on the level of individual genes (Fig. 2b). For example, the ESC transcriptional regulator Esrrb[7] falls into the category 2 concordance pattern at all time points. This indicates that ultimate levels of Esrrb protein are primarily regulated post-transcriptionally, at least under our experimental conditions, and not by direct Nanog regulation at the transcriptional level. It has been proposed that Esrrb and Nanog mutually regulate each other by a positive feedback circuit[6,18]. Our concordance pattern analysis of Esrrb indicates that at least one other component is likely to be involved in this circuit, which is responsible for the post-transcriptional regulation of Esrrb, possibly a microRNA[19,20].

Gene-ontology analyses across the four molecular layers suggest a complex interaction between different molecular regulatory mechanisms in cell fate regulation (Fig. 2c and Supplementary Fig. 5). For example, differentiation- and development-related genes are overrepresented among the genes that only show changes in acH3K9/14 levels, but not on the other three layers (Fig. 2c). Furthermore, chromatin- and nucleosome-assembly-related genes are overrepresented among the genes upregulated on the RNA polymerase II binding layer but not on any of the other three layers (Fig. 2c), suggesting that the chromatin modifiers are primarily regulated at the transcription step. Therefore, reconfiguration of chromatin structure, although an important factor in ESC fate alteration, may have a secondary role to primary regulation by transcription factors[5,6,8,21–23].

To gain further insight into systems-level regulatory control of changes in ESC fate, we combined our data with that of previous stem cell regulatory network studies to form a new synthesis (Fig. 3)[6,8,24]. A core protein–protein interaction network was previously identified in murine ESCs involving 26 proteins centred around Nanog[24]. We found that this interactome is enriched in proteins that decreased in expression after downregulation of Nanog (Supplementary Fig. 6). On day 5, 8 out of the 26 interactome proteins are at significantly reduced levels (Supplementary Fig. 7). These are: Sall4, Rnf2, Oct4 (also known as Pou5f1), Ilf2, Nanog, Mybbp1a, Sall1 and Esrrb. Of these eight proteins only one (Rnf2) does not directly interact with Nanog (Fig. 3a). This suggests interdependence between the Nanog interactome and the network of genes under Nanog transcriptional control.

Nanog protein binds to thousands of genomic locations in undifferentiated ESCs[5,6]. Our data show that approximately 20% of the previously identified Nanog-binding genes change their transcription levels (POL) during the first 5 days after Nanog downregulation.

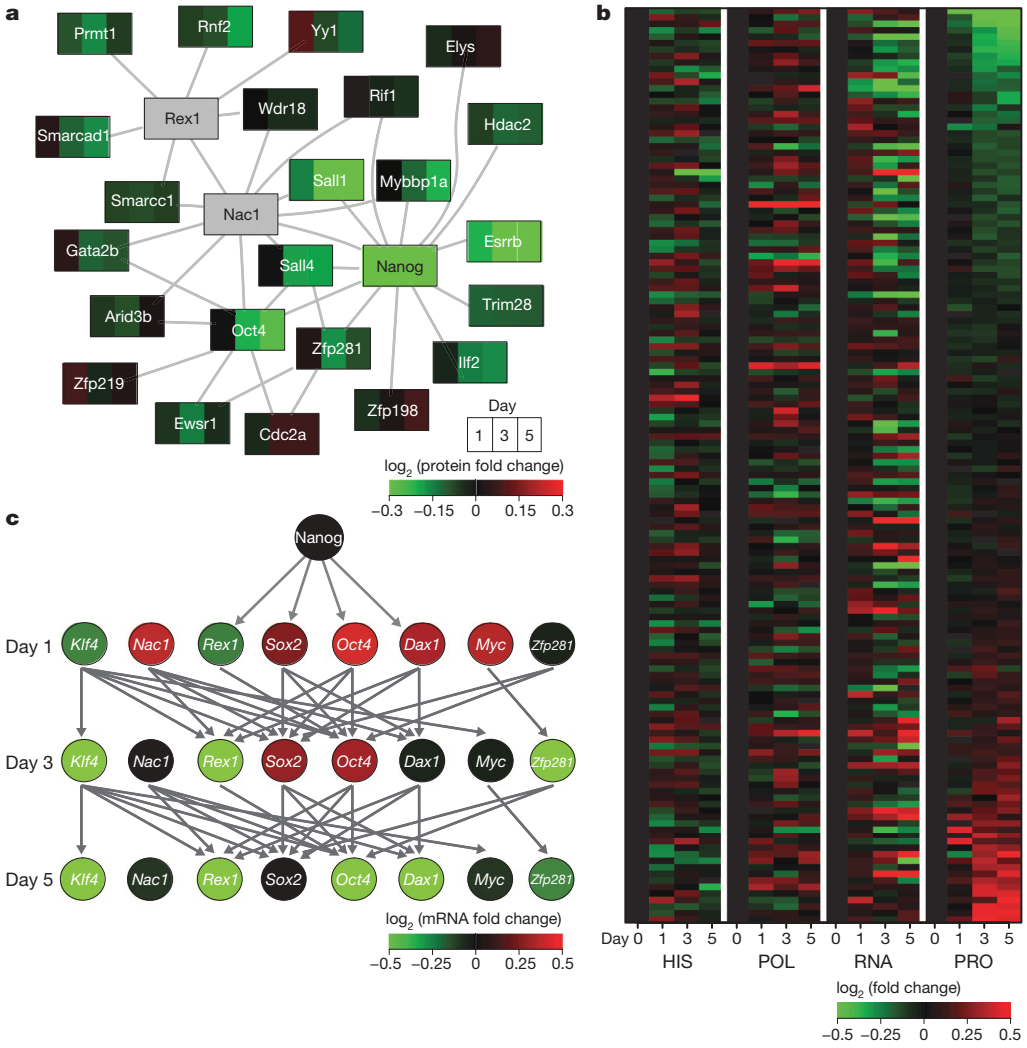


Figure 3 | Dynamic changes in ESC networks. **a**, The core ESC protein–protein interaction network[24] (connections) overlaid with dynamic protein changes observed in our data (rectangles are divided into three segments representing changes on days 1, 3 and 5 compared to day 0). **b**, Heat map of multimolecular layer gene expression changes for Nanog-binding genes[6]. Shown are the genes whose data were obtained with high confidence on all four molecular layers. Genes are ranked on the basis of changes in protein levels. **c**, The pluripotency transcriptional regulatory network[8] (arrows) overlaid with mRNA fold changes (colours) from our data.

Of those that changed, approximately 50% also exhibit changes in protein levels (PRO) (Fig. 3b and Supplementary Fig. 7). To determine how the changes in expression develop after the downregulation of Nanog, we analysed the temporal alterations of mRNAs in the context of an extended transcriptional regulatory network proposed previously[8] (Fig. 3c). Our data show that most genes in this network are downregulated after the removal of Nanog. In particular, downregulation of Oct4 and Sox2 (protein levels shown in Supplementary Fig. 7) occurred later than downregulation of Klf4 or Rex1. This suggests that decreases in Oct4 and Sox2 expression are not responsible for decreases in Klf4 and Rex1 expression under our experimental conditions. The temporal sequence of changes in gene expression is loosely correlated with the chromatin-binding data[6,8]. These two sources provide independent and complementary information about the ESC gene regulatory network. Using the same principle that later molecular events cannot regulate earlier events, we can extract new sets of useful information concerning the gene regulatory relations from the temporal order of the network reconfiguration (Fig. 4 and Supplementary Fig. 8).

To facilitate comparisons and visualization of the multilayered time series, we generated interactive movies to display our data (Fig. 4 and Supplementary Fig. 8; http://amp.pharm.mssm.edu/ronglu). Expression changes for 400 genes with the most significant changes in protein levels on day 5 were projected onto two-dimensional hexagonal arrays (Fig. 4a). Individual hexagons representing specific genes are dynamically coloured according to the fold changes in each of the four molecular layers. This approach facilitates genome-wide and temporal comparisons among the different molecular layers, and allows clustering of genes with similar dynamics on multiple gene expression regulatory layers. We have also generated interactive scatter plot movies to help visualize concurrent changes across the different molecular layers (Fig. 4b). In these movies, individual genes can be selected to illustrate the concurrent changes between pairs of molecular layers. For instance, Fig. 4b demonstrates that changes in Esrrb mRNA and protein expression are monotonically related, whereas Sall1 and Oct4 both show increased mRNA levels without any corresponding increase in protein levels during the early stage of ESC differentiation. Similar dynamics are also exhibited by several other previously identified essential ESC factors[25] (shown as red dots in Fig. 4b). These genes are regulated on different regulatory layer(s) compared to Esrrb, and suggest that the transcription layer undergoes an early cell fate reconfiguration without significant accompanying changes in protein production. Recent studies proposed that fluctuating levels of Nanog may discriminate between alternative pluripotent states of ESCs, in which high or low Nanog levels render ESCs resistant or susceptible to differentiation inducing stimuli, respectively[15,26–29]. In our system, the early time point of Nanog downregulation is comparable to the 'low' Nanog state from these studies. Thus, the absence of changes in protein levels during the mRNA layer reconfigurations could reflect the nature of these distinct pluripotent states. Collectively, the variety of the multilayered expression patterns underscores the complexity of the molecular regulation of ESC fate and suggests an intricate regulatory network involving several molecular regulatory layers.

In this study we have provided a dynamic multimolecular layer view of a murine ESC fate change in response to the downregulation of Nanog. In our experimental system the transcription of *Nanog* is regulated by exogenous manipulation and not by the endogenous regulatory circuit. This disrupts the balance of mutually regulated ESC molecular circuits[15,26–29], and allows for rapid and synchronous cell fate changes within the population. However, our results nonetheless represent the average of a large cell population, as we have shown previously that removing Nanog results in a complex mixture of cell lineages[7]. In this work, our primary goal was to analyse the molecular dynamics that are associated with the transition away from the pluripotent state as it occurs in most of the cells. *In vivo*, cell fate alteration is probably triggered by several perturbations and inputs dynamically. The single gene perturbation that we have used does not reflect the natural signals that pluripotent cells are subjected to *in vivo*. However, it is a powerful tool to dissect the complex regulatory networks that underpin ESC fate changes and offers an initial window into the dynamic complexity of ESC fate regulation across multiple molecular levels.

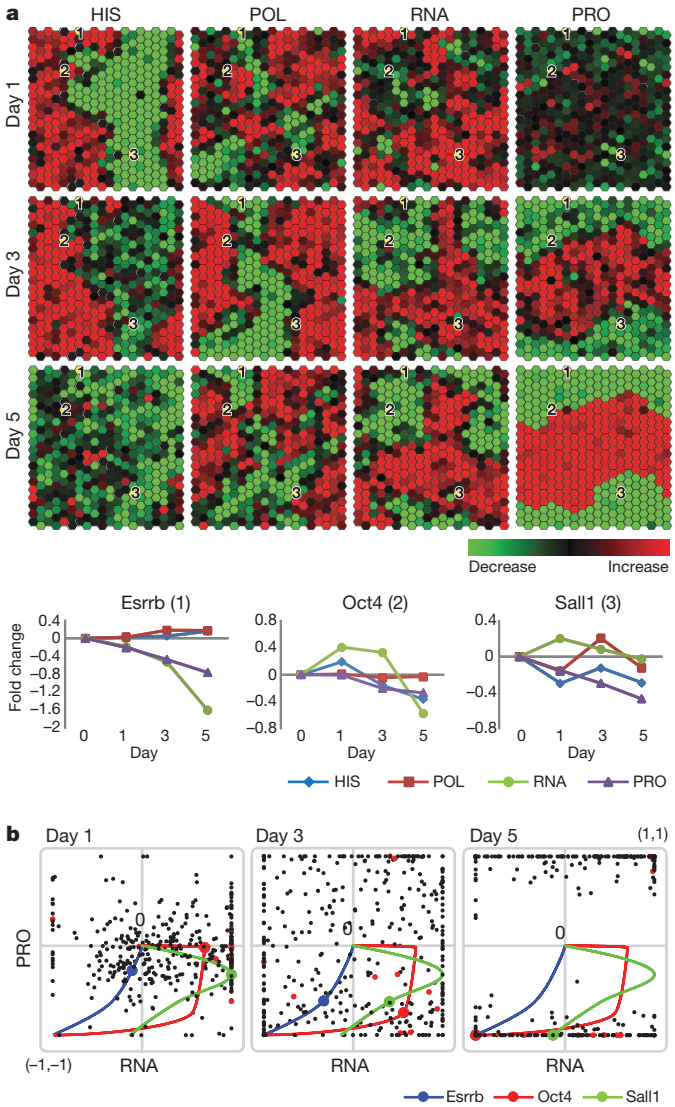


Figure 4 | Interactive visualization of the multilayer dynamic data. **a**, Snapshots from heat map movies showing 400 genes with the most significant changes in protein levels on day 5. The position (pixel) of each gene locus is the same in all 12 heat maps. **b**, Snapshots from dynamic scatter plots illustrating concurrent changes in mRNAs and proteins. Red dots represent genes that have been identified to have important roles in ESCs[25]. Supplementary Fig. 8 and the website http://amp.pharm.mssm.edu/ronglu are interactive and each gene can be displayed as a line plot as exemplified by *Esrrb*, *Oct4* and *Sall1*.

METHODS SUMMARY

AcH3K9/14 levels were assayed using ChIP-chip. Acetylated regions in a 1-kilobase window around the transcription initiation position were identified to generate acetylation profiles (Supplementary Figs 9 and 10). ChIP-chip was also used to measure RNA polymerase II localization on 3′ exons to directly assess transcriptional activity (elongation). Changes in mRNA levels were monitored using Agilent two-colour microarrays. Nuclear protein levels were measured using peptide isobaric tagging followed by two-dimensional liquid chromatography mass spectrometry (LC-MS/MS)[16]. We chose to measure nuclear protein levels because cell fate determination is largely controlled in the nucleus. For technical reasons, attempts to measure the entire proteome would have significantly decreased the sensitivity of the nuclear protein measurements, as these only constitute approximately 20% of all

proteins in ESCs. All experiments were conducted in triplicate except for the acH3K9/14 measurements, which were performed in duplicate. Reliability of all data sets was verified using independent experimental assays, including conventional chromatin immunoprecipitation (ChIP), quantitative PCR (qPCR), and western blot assays for key pluripotency regulator genes (Supplementary Figs 11 and 12). Experimental reproducibility was also verified using a linear analysis of variance (ANOVA) model[30]. After data pre-processing and normalization, we were able to validate 1,627 nuclear proteins and 12,488 genes (HIS/POL/RNA) with high confidence. For 1,212 nuclear proteins, we were able to jointly obtain high-quality data across all four layers (HIS/POL/RNA/PRO). Supplementary Fig. 1 provides an overview of the entire data processing pipeline and the results of the quality-control procedures (ANOVA analysis). The significance of change is determined at a FDR of 5% using an empirical Bayes' model with Benjamini–Hochberg correction on the basis of experimental replicates.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 12 February; accepted 9 October 2009.

1. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).
2. Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441,** 349–353 (2006).
3. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
4. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454,** 766–770 (2008).
5. Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122,** 947–956 (2005).
6. Loh, Y. H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genet.* **38,** 431–440 (2006).
7. Ivanova, N. *et al.* Dissecting self-renewal in stem cells with RNA interference. *Nature* **442,** 533–538 (2006).
8. Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S. H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132,** 1049–1061 (2008).
9. Chickarmane, V. & Peterson, C. A computational model for understanding stem cell, trophectoderm and endoderm lineage determination. *PLoS One* **3,** e3478 (2008).
10. Ying, Q. L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453,** 519–523 (2008).
11. Sampath, P. *et al.* A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation. *Cell Stem Cell* **2,** 448–460 (2008).
12. Chang, W. Y. & Stanford, W. L. Translational control: a new dimension in embryonic stem cell network analysis. *Cell Stem Cell* **2,** 410–412 (2008).
13. Chambers, I. *et al.* Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* **113,** 643–655 (2003).
14. Mitsui, K. *et al.* The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113,** 631–642 (2003).
15. Chambers, I. *et al.* Nanog safeguards pluripotency and mediates germline development. *Nature* **450,** 1230–1234 (2007).
16. Unwin, R. D. *et al.* Quantitative proteomics reveals posttranslational control as a regulatory factor in primary hematopoietic stem cells. *Blood* **107,** 4687–4694 (2006).
17. Spooncer, E. *et al.* Developmental fate determination and marker discovery in hematopoietic stem cell biology using proteomic fingerprinting. *Mol. Cell. Proteomics* **7,** 573–581 (2008).
18. van den Berg, D. L. *et al.* Estrogen-related receptor β interacts with Oct4 to positively regulate *Nanog* gene expression. *Mol. Cell. Biol.* **28,** 5986–5995 (2008).
19. Tay, Y., Zhang, J., Thomson, A. M., Lim, B. & Rigoutsos, I. MicroRNAs to *Nanog, Oct4* and *Sox2* coding regions modulate embryonic stem cell differentiation. *Nature* **455,** 1124–1128 (2008).
20. Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134,** 521–533 (2008).
21. Bonifer, C., Hoogenkamp, M., Krysinska, H. & Tagoh, H. How transcription factors program chromatin—lessons from studies of the regulation of myeloid-specific genes. *Semin. Immunol.* **20,** 257–263 (2008).
22. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).
23. Okita, K., Ichisaka, T. & Yamanaka, S. Generation of germline-competent induced pluripotent stem cells. *Nature* **448,** 313–317 (2007).
24. Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444,** 364–368 (2006).
25. Macarthur, B. D., Ma'ayan, A. & Lemischka, I. R. Systems biology of stem cell fate and cellular reprogramming. *Nature Rev. Mol. Cell Biol.* **10,** 672–681 (2009).
26. Graf, T. & Stadtfeld, M. Heterogeneity of embryonic and adult stem cells. *Cell Stem Cell* **3,** 480–483 (2008).
27. Dietrich, J. E. & Hiiragi, T. Stochastic patterning in the mouse pre-implantation embryo. *Development* **134,** 4219–4231 (2007).
28. Singh, A. M., Hamazaki, T., Hankowski, K. E. & Terada, N. A heterogeneous expression pattern for Nanog in embryonic stem cells. *Stem Cells* **25,** 2534–2542 (2007).
29. Kalmar, T. *et al.* Regulated fluctuations in Nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol.* **7,** e1000149 (2009).
30. Airoldi, E. M. Getting started in probabilistic graphical models. *PLOS Comput. Biol.* **3,** e252 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank D. Storton for technical support, and E. Wieschaus, Y. Shi, S. Tavazoie and N. Slavov for constructive discussions. We also acknowledge the laboratories of the following people for providing antibodies for western blot: A. Okuda, J. Flint and Y. Kang. This work was supported by the NIH, and in part supported by the BBSRC and Leukaemia Research UK. O.G.T., F.M. and E.M.A. were partially supported by the NIH and US National Science Foundation.

Author Contributions R.L. and I.R.L. designed the experiments. R.L. prepared the cell samples for all the experiments, performed the RNA polymerase II ChIP-chip, the mRNA microarray, and verification experiments such as western blot, ChIP and quantitative PCR. R.D.U. and A.D.W. performed the proteomic experiments and primary analysis on proteomic data. L.A.B. performed the histone acetylation ChIP-chip experiments. R.L., F.M., E.M.A., R.R. and O.G.T. performed general data processing and statistical analyses. R.L. and F.M. plotted Figs 1–3. A.L., B.D.M. and A.M. developed and plotted interactive Fig. 4a. A.L. and A.M. developed and plotted interactive Fig. 4b. R.L., J.L., F.M. and I.R.L performed network analysis shown in Fig. 3. R.L. and I.R.L. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to I.R.L. (ihor.lemischka@mssm.edu) or R.L. (rlu@stanford.edu).

METHODS

Cell culture. A murine ESC line with controllable Nanog expression was constructed and characterized previously[7], and was cultured as described. ESCs were cultured without feeder cells (primary mouse embryonic fibroblasts) for all experiments. To induce differentiation, we withdrew doxycycline ($1\,\mu g\,ml^{-1}$) from the media, but still maintained all of the routine ESC nutrients (DMEM with 15% FBS (Hyclone), 100 mM MEM non-essential amino acids, 0.1 mM 2-mercaptoethanol, 1 mM L-glutamine (Invitrogen) and $10^3\,U\,ml^{-1}$ LIF (Chemicon)).

Days 1, 3 and 5 were selected because: (1) our previous studies[7], which investigated the differentiation process using microarrays and quantitative PCR analysis over the course of 12 days, suggested that days 1, 3 and 5 are sufficiently early such that no major differentiation events have yet occurred in the population, but are also sufficiently late and temporally spaced to study transitions from pluripotency and temporal differences. (2) Our preliminary proteomic experiment had shown that a reasonable number of proteins had changed during this time frame. In other words, on days 1, 3 and 5 the numbers of proteins that had changed were large enough to analyse using mass spectrometry, and were also small and distinct enough from each other to study the initial dynamic changes (Fig. 1d).

ChIP-chip. ChIPs were performed as described[5]. Fifty-million to five-hundred-million cells were fixed in a formaldehyde solution and sonicated into chromatin fragments containing 500–1,000 base pairs of DNA. ChIP was performed using 100 µl of a protein G magnetic bead suspension from Dynal coated by 10 µg of antibody (anti-acH3K9/14 (06-599) from Upstate; RNA polymerase II antibody (MMS-126R) from Covance). After reversal of the cross-links, the isolated DNA and non-ChIP-enriched control DNA were tailed with polyA by terminal transferase (TdT)[31]. T7 $(dT)_{24}$ primer was used to incorporate the T7 promoter during the second-strand synthesis reaction. The DNA fragments were then linearly amplified and labelled with Cy3 and Cy5 during the *in vitro* transcription, following the protocol provided by Agilent for dye incorporation and array hybridization (Agilent low RNA input fluorescent linear amplification kit; Agilent 60-mer oligo microarray processing protocol version 2.1). The histone acetylation ChIP-chip was performed by L. Boyer. The amplification step is slightly different[5].

Microarrays. The histone acetylation ChIP-chip used the mouse promoter array from Agilent, custom-designed by the R. Young laboratory. An Agilent whole mouse genome oligonucleotide microarray that covered 41,000 well-characterized mouse genes and transcripts was used for the mRNA assays and RNA polymerase II ChIP-chip experiments.

Nuclear proteome. Nuclear protein samples were prepared with the Nuclear/Cytosol fractionation kit (BioVision). Proteomic measurements were performed according to published protocols[16]. Samples from four different time points (day 0, and days 1, 3 and 5 after doxycycline removal) were labelled using four-channel isobaric tagging reagents (iTRAQ, Applied Biosystems) and analysed by strong cation-exchange fractionation followed by reverse-phase liquid chromatography on line to a QStar XL quadropole time-of-flight mass spectrometer. We used ProQUANT (Applied Biosystems) and ProGROUP (Applied Biosystems) to identify and quantify proteins. We checked our proteomic data with the proteomic data from a previous study[32]. Only 3.1% of the proteins that we considered to be well-reproduced nuclear proteins were not identified as nuclear proteins in their study.

Data confirmation (qPCR and western blot). ChIP-chip results for RNA polymerase II localization and histone acetylation were verified using a commercial ChIP kit (Upstate), followed by qPCR. RNA microarray data were verified by qPCR. The qPCR kit was obtained from Stratagene (Brilliant SYBR Green QPCR Master Mix). Proteomic data were confirmed using western blot. The verification experimental results are shown in Supplementary Figs 11 and 12. Antibodies used to perform western blot were: Oct4 antibody from BD; Nanog antibody from Cosmo Bio; Dnmt3b antibody from Abgent; p53 antibody from J. Flint; β-actin antibody from Santa Cruz; HSP 90 antibody from Upstate; Histone H1.0 antibody from Abcam; Utf1 antibody from A. Okuda.

Processing microarray data. Background correction was performed using a Normal+Exponential convolution model[33] that adjusts the foreground to the background and yields strictly positive intensities. Furthermore, we used an offset to dampen the variation of the log-ratios for very low intensities near 0. This stabilized our estimated fold changes. Arrays were normalized using a global loess, which is a well-tested general-purpose normalization method using local regressions to straighten the 'banana-shape' seen in raw measurements[34]. To confirm data quality, microarrays with remaining spatial (and other) artefacts were discarded and the experiments repeated.

Processing proteomic data. We used ProQUANT and ProGROUP software (Applied Biosystems) to analyse the mass spectrometric data, giving confidence values for the relative quantification analysis. Our proteomic analysis was based only on proteins that were identified with more than 95% confidence. We further filtered proteins based on two filters: (1) filter criteria based on raw data: the error factor of the measurement must be smaller than 2 and the protein must have been detected in at least two of the three runs. (2) Assessing reproducibility of protein measurements: we fitted a linear model (two-way ANOVA) to obtain temporal and replicate effects. If a significant replicate effect existed, we deemed the protein to be 'non-reproducible' and discarded it from further analysis.

Identification of histone acetylation regions. We compared the measurement for each probe on the promoter array against the distribution of measurements for all the negative control probes (null distribution), and calculated a *P* value for every probe (Supplementary Fig. 9). We use a FDR cut-off of 0.1 on the *P* value distribution to define which probes were acetylated and which were not. Supplementary Fig. 10 shows example acetylation profiles that indicate the acetylated regions and illustrate the main changes that occurred there.

Assessing experimental reproducibility and merging data. (1) Assessing reproducibility of probes: for every microarray probe, we fitted a linear model[30,35] (two-way ANOVA) to extract temporal and replicate effects. If a probe had a significant replicate effect, we deemed it to be non-reproducible and discarded it from further analysis. (2) Averaging probes that represent the same gene: for RNA polymerase II and mRNA expression data, we performed a three-way ANOVA with temporal, replicate and probe effects. Only genes with non-significant probe effects were used for further analysis (that is, those for which all probes behave coherently). For the histone acetylation data set, we averaged acetylated probes in a 1-kilobase window around the transcription start position (red lines in Supplementary Fig. 10 mark this region). (3) Combining gene isoforms: data from different molecular layers were merged based on our ID matching strategy (details later). For genes with more than one isoform, we applied a three-way ANOVA to determine temporal, replicate and gene effects. If the gene effect was significant (showing non-coherent behaviour), we discarded the data. The data for each gene in each data set at each time point were averaged if coherent behaviour was observed on both probe and gene levels.

ID matching. We matched protein IDs, microarray IDs, and MGI symbols (for GoMiner) using Ensembl BioMart (http://www.ensembl.org/Multi/martview), supplemented with protein information from the following databases: http://www.ebi.uniprot.org/uniprot-srv/uniProtEntryListSearch.do; http://www.ncbi.nlm.nih.gov/entrez/batchentrez.cgi?db=Nucleotide; and http://www.pir.uniprot.org/search/idmapping.shtml. Histone acetylation ChIP-chip data were matched to the RNA microarray data using the UCSC database and the Ensembl database. The ID match file is included in the Supplementary Information.

Determining significance thresholds. For each of the four data sets, we computed the standard deviations of each gene using values from the replicate experiments. We then used the median value of the entire set of standard deviations in each data set as an estimate of the experimental error. For each of the four data sets, *P* values were independently computed using a Gaussian model for the measurements of each gene, under the null hypothesis given by setting the mean at zero, and the standard deviation at the experimental error estimate. Corrected *P* value was then obtained using the FDR correction[36]. Up- and downregulated genes were considered to be significant at a confidence level of $\alpha = 0.05$. An overview of the results is given in Supplementary Fig. 2, which shows the number of up- and downregulated genes in all data sets and for genes with protein data.

Methods for Fig. 2a. The method we used to generate Fig. 2a is not a clustering per se. Conventional clustering method is only applied at the very last step for visualization, but does not determine the categories. Our method is basically an iterative gene selection procedure, starting on the PRO level and working from there 'backwards' to RNA, POL and finally to HIS. The step-by-step description is as follows: (1) for each day, select all genes with significant protein changes. Genes without significant protein changes are discarded. (2) Select all genes that show a direction of change on the PRO level that is opposite to that on the RNA level. These genes form category 1. (3) Select all genes that show the same direction of change on PRO and RNA, but the opposite direction on POL. These genes form category 2. (4) Select all genes that show the same direction of change in PRO, RNA and POL, but the opposite direction in HIS. These genes form category 3. (5) All remaining genes show the same direction of change in all layers—PRO, RNA, POL and HIS. These genes form category 4. (6) Within each category, we cluster the genes with standard hierarchical clustering ('hclust' function in R) using complete linkages and a Euclidean distance. This clustering does not influence the definition of the four categories. It only improves the 'readability' of the resulting heatmap. Data are normalized within each column (molecular layer).

Methods for Fig. 4 and online movies. We selected the 400 genes with the most significant changes in protein expression on day 5. Because there are four time points (days 0, 1, 3 and 5), the data from each molecular layer is a 400×4 matrix. To consider correlations across layers, we first concatenated the time series from

all four layers into a 400 × 16 data matrix D. To visualize systems-level regulatory dynamics we then projected this data matrix onto a regular hexagonal array H by assigning each row of the data matrix to a unique hexagon h in H. A hexagonal array was chosen because it presents the data in a form that is easy to visualize. To provide a continuous geometric object with no boundaries we associated the left- and right-hand sides of the array with each other, and the top and the bottom of the array with each other (to make the surface of a torus). These conditions ensure that there are no special places on the array and all molecular species are treated equally.

Not all arrangements of the data on the array will capture the system-level regulatory dynamics equally well: most arrangements will not capture the collective dynamics because molecular species with similar expression patterns will not be close to each other on the array. To construct an arrangement that best captures collective dynamics we assigned to each arrangement a fitness

$$\text{Fit} = \frac{1}{2{,}400}\sum_{i=1}^{400}\sum_{j\in N_j} C_{ij},$$

in which C_{ij} is the Pearson's correlation coefficient between the time series i and j, and N_j are the six neighbours of the hexagon h_j. Fit measures how well a given arrangement captures the collective dynamics of the system in general: arrangements with low fitness do not capture system-level dynamics, whereas arrangements with high fitness capture system-level dynamics well. To find the arrangement of the time series on the array with the maximal fitness we used a simulated annealing algorithm, and ran the annealing algorithm overnight (12 h) to ensure as close to an optimal arrangement as possible.

Movies of systems-level dynamics were then generated by dynamically assigning colours to each of the hexagons in the array based upon the expression fold changes of the gene to which it is assigned. To create a movie that interpolates smoothly between time-points, each time series was normalized such that all expression series range from 0 to 1 and a piecewise cubic Hermite interpolation was implemented before visualization. Similar movies can be created using GATE (http://amp.pharm.mssm.edu/maayan-lab/gate.htm), a system we developed for this purpose.

We note here that the clustering technique we have used is similar to a self-organizing map (SOM), and the movies we create are similar to those created by the Gene Expression Dynamics Inspector (GEDI)[37] using SOMs. Given a set of time-series data describing expression changes in a large number of genes, the GEDI uses SOMs to project the expression time series onto a two-dimensional rectangular array, and colours rectangles according to the genes to which they are associated. However, because the GEDI uses a SOM, individual rectangles are associated with a cluster of genes that share similar expression patterns. In our study, we are concerned with the gene expression at different molecular layers. Thus it was important to track the molecular regulation of individual (rather than clusters of) genes. For this reason, we used the above custom-written algorithm that assigns molecular species to hexagons in a strictly one-to-one manner.

Code and software. Data pre-processing, data normalization and large parts of the analysis were performed in the computing languages Python and R (http://www.r-project.org/) using packages available from the Bioconductor website (http://www.bioconductor.org/). In particular, we relied on the limma package (http://bioinf.wehi.edu.au/limma/) including the Norm-Exp model for background correction as described previously[33]. To create Fig. 4, we used GATE (http://amp.pharm.mssm.edu/maayan-lab/gate.htm) and AS3/Flash. Our pre-processing and analysis pipeline is available from the authors on request.

31. Liu, C. L., Schreiber, S. L. & Bernstein, B. E. Development and validation of a T7 based linear amplification for genomic DNA. *BMC Genomics* **4,** 19 (2003).
32. Kislinger, T. *et al.* Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* **125,** 173–186 (2006).
33. Ritchie, M. E. *et al.* A comparison of background correction methods for two-colour microarrays. *Bioinformatics* **23,** 2700–2707 (2007).
34. Yang, Y. H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30,** e15 (2002).
35. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3,** Article3 (2004).
36. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* **125,** 279–284 (2001).
37. Eichler, G. S., Huang, S. & Ingber, D. E. Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles. *Bioinformatics* **19,** 2321–2322 (2003).

LETTERS

Signal peptides are allosteric activators of the protein translocase

Giorgos Gouridis[1,2], Spyridoula Karamanou[1], Ioannis Gelis[3], Charalampos G. Kalodimos[3] & Anastassios Economou[1,2]

Extra-cytoplasmic polypeptides are usually synthesized as 'preproteins' carrying amino-terminal, cleavable signal peptides[1] and secreted across membranes by translocases. The main bacterial translocase comprises the SecYEG protein-conducting channel and the peripheral ATPase motor SecA[2,3]. Most proteins destined for the periplasm and beyond are exported post-translationally by SecA[2,3]. Preprotein targeting to SecA is thought to involve signal peptides[4] and chaperones like SecB[5,6]. Here we show that signal peptides have a new role beyond targeting: they are essential allosteric activators of the translocase. On docking on their binding groove on SecA, signal peptides act *in trans* to drive three successive states: first, 'triggering' that drives the translocase to a lower activation energy state; second, 'trapping' that engages non-native preprotein mature domains docked with high affinity on the secretion apparatus; and third, 'secretion' during which trapped mature domains undergo several turnovers of translocation in segments[7]. A significant contribution by mature domains renders signal peptides less critical in bacterial secretory protein targeting than currently assumed. Rather, it is their function as allosteric activators of the translocase that renders signal peptides essential for protein secretion. A role for signal peptides and targeting sequences as allosteric activators may be universal in protein translocases.

We sought to dissect the individual contributions of signal peptides and mature domains to membrane targeting and to post-targeting translocation steps. Because SecB is not universal or essential[6,8], we used the SecB-independent[9,10] substrate proPhoA (periplasmic alkaline phosphatase).

The affinity of proPhoA for inverted inner membrane vesicles (IMVs) containing SecYEG either alone or complexed with SecA was determined (Fig. 1a). ProPhoA associates with high affinity (0.23 μM) to SecYEG-bound SecA, but not to SecYEG alone. Like proOmpA[5], proPhoA requires SecA as an essential receptor. ProPhoA association to SecYEG-bound SecA is only marginally reduced if the signal peptide is impaired (proPhoA(L8Q), proPhoA(L14R), see refs 9–11). In contrast, this binding is reduced sevenfold once the mature region is carboxy-terminally truncated (proPhoA(1–62)). Therefore, the mature domain moiety contributes substantially to proPhoA translocase binding. This association was quantified for the first time: PhoA associated with SecYEG-bound SecA almost as strongly as proPhoA (0.6 μM), demonstrating that mature domains contain prominent targeting determinants. A large excess of signal peptide, added *in trans*, does not out-compete PhoA binding (Supplementary Fig. 2). Thus, signal peptide[4] and mature domain[12] binding sites on SecA must be distinct.

Soluble SecA also binds proPhoA and its derivatives tightly, with ~1:1 stoichiometry (Fig. 1a and Supplementary Fig. 3b and c). This implies similar recognition of mature domains by SecYEG-bound and cytoplasmic SecA, although the latter interaction is ~tenfold weaker. Because a synthetic proPhoA signal peptide binds to soluble SecA with fivefold less affinity than PhoA (Fig. 1a), the mature domain is the primary binding determinant.

Periplasmic PhoA folds into its native, enzymatically active structure after forming intramolecular disulphides[13]. In the reducing, cytoplasm-like, environment used above, proPhoA exists in a 'non-native' state, has no phosphatase activity (Supplementary Fig. 1c) and is translocation-competent (Supplementary Fig. 1a and b, lane 3). Oxidized, 'native' proPhoA is an active phosphatase, like PhoA[10,13] (Supplementary Fig. 1c), but is translocation-incompetent (Supplementary Fig. 1a, lane 4 and b, lane 2). 'Native' proPhoA cannot associate with either soluble or SecYEG-bound SecA (Supplementary Fig. 3a and d) although it carries a functional signal peptide. Thus, mature domain targeting signals required for docking at the membrane are only presented on 'non-native' preproteins.

We next turned to post-targeting events. Is docking of a mature domain to the translocase sufficient to ensure secretion across the membrane? In contrast to proPhoA that was proficiently secreted *in vivo* or *in vitro* (Fig. 1b–d, lane 1)[9,10], secretion of PhoA (lane 4), proPhoA(L14R) (lane 3) or proPhoA(L8Q) (lane 2) was marginal. Clearly, functional signal peptides are essential for translocation of docked mature domains.

To identify what is the essential role of signal peptides in translocation, we examined their effect on ATP hydrolysis by the translocase (Fig. 2; Supplementary Fig. 4)[14]. Translocating preproteins lower the ATPase activation energy (E_a) markedly (Fig. 2a; compare lane 3 to 1 and 2; Supplementary Fig. 4c) and stimulate multiple ATP turnovers by six- to ninefold ('translocation ATPase'; Fig. 2b, lane 1; Supplementary Fig. 4a)[14,15]. Wild-type proPhoA synthetic peptide alone fully retains the ability to lower E_a to a similar extent (Fig. 2a, lane 7) and in a saturable manner (Supplementary Fig. 4d). In contrast, its L8Q or L14R derivatives (Fig. 2a, lanes 8 and 9), mature PhoA, proPhoA(L8Q) and proPhoA(L14R) (lanes 4–6) all fail to lower E_a. Therefore, functional signal peptides are necessary and sufficient to lower the activation energy state of the translocation ATPase, an effect we term 'triggering'. However, because neither signal peptide nor PhoA alone stimulated translocation ATPase (Fig. 2b, lanes 5 and 2), several ATP turnovers[15] require ongoing 'secretion' of mature domains (Fig. 2b, lane 1)[7,14,15]. These two steps, 'triggering' and 'secretion', are ordered and can be experimentally dissected.

To uncouple them we used *prl* (protein localization) mutants in *sec* genes. We reasoned that these might bypass triggering because they allow some secretion of preproteins with defective or missing signal peptides *in vivo*[9–11,16–18]. Indeed, the PrlA4 (refs 10, 16 and 19) and PrlA3 (ref. 16) (not shown) mutant translocases are constitutively triggered in the absence of any preprotein (Fig. 2a, compare lane 10

[1]Institute of Molecular Biology and Biotechnology, Foundation of Research and Technology-Hellas, Iraklio, Crete 71110, Greece. [2]Department of Biology, University of Crete, Iraklio, Crete 71409, Greece. [3]Chemistry & Chemical Biology, Biomedical Engineering, Rutgers University, 599 Taylor Rd, Piscataway, New Jersey 08854, USA.

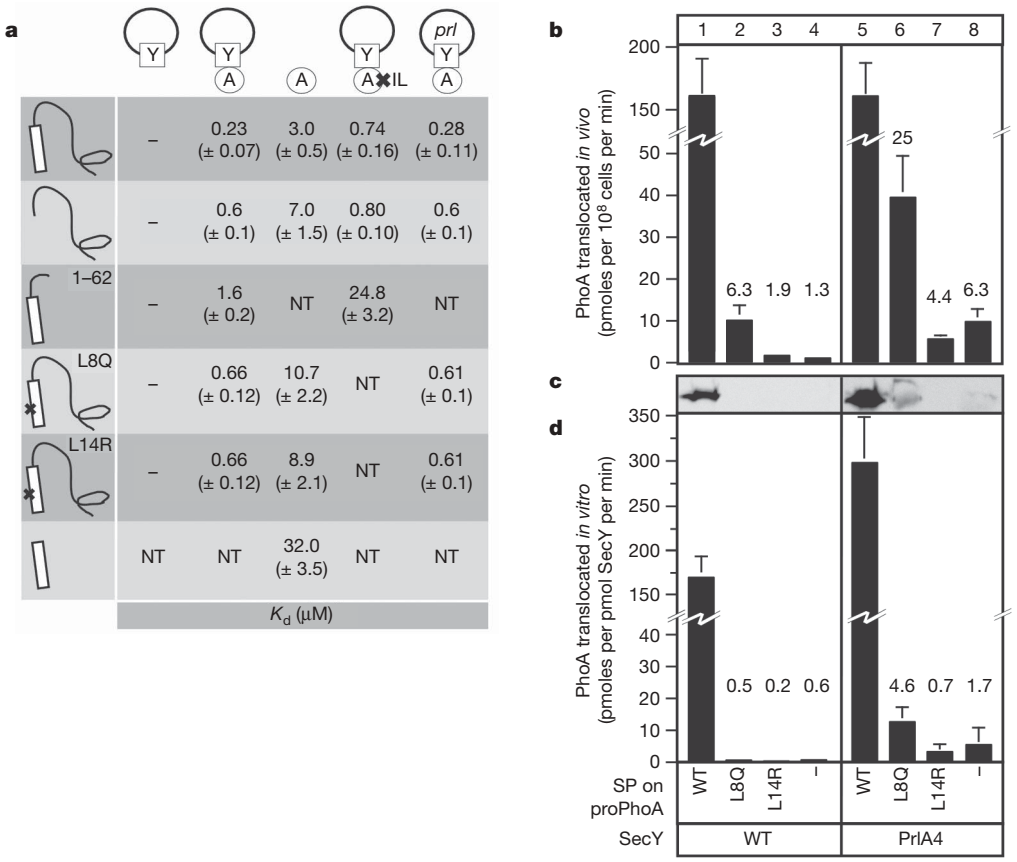


Figure 1 | Translocase binding and export of proPhoA and its derivatives. **a**, Equilibrium dissociation constants (K_d) of proPhoA and variants for the translocase. SecA(I304A/L306A) (marked 'IL') or PrlA4/SecYEG were used (n = 3–7). A, SecA; Y, SecY; **x**, Mutant derivative; −, no detectable binding; NT, not tested. **b**–**d**, *In vivo* (**b**) or *in vitro* (**c** and **d**) translocation of proPhoA and derivatives by wild-type or PrlA4/SecYEG translocase. In **b** phosphatase units were converted to protein mass. Proteins visualized by immunostaining (**c**) were quantified by phosphorimaging (**d**).The percentage of translocated material compared to that of the wild-type proPhoA (100%) is indicated above each bar (n = 9). In all the experiments the mean (or average value) of the independent experiments (n) is indicated. Error bars (Figs 1b, d, 2b, 3b) represent the standard deviation (or the square root of the variance) of all the values of the independent experiments.

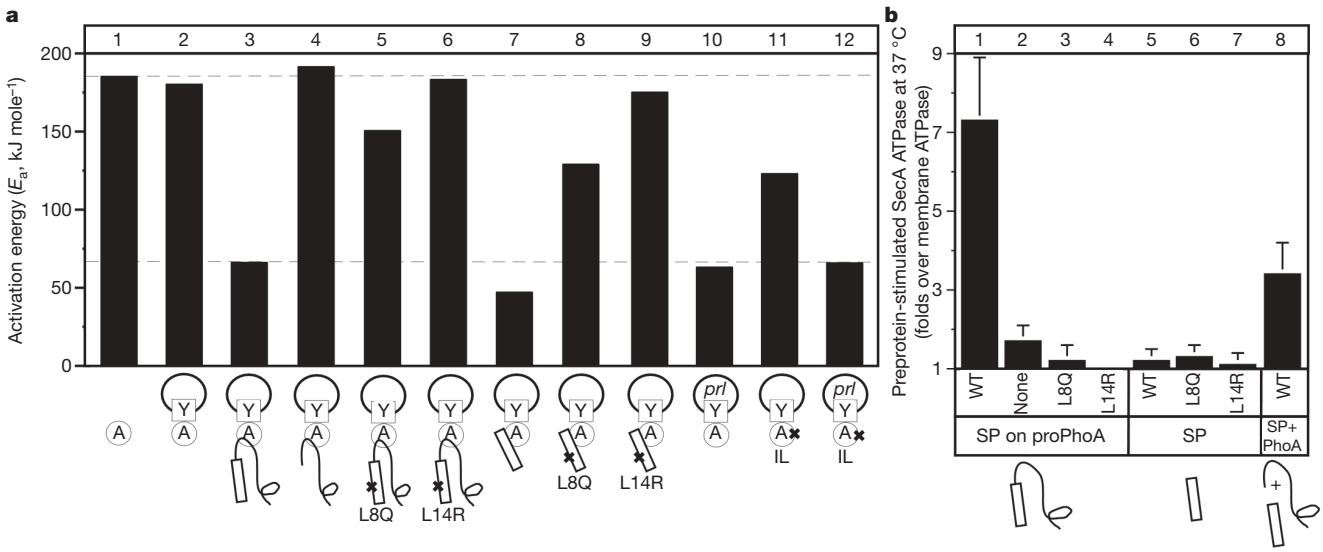


Figure 2 | Activation energy and stimulation of SecA ATPase under different regimes. **a**, The activation energy (E_a; kJ mole^{-1}) of the wild-type translocase and variants was determined in the presence of various preprotein derivatives and in the presence of synthetic signal peptides, as indicated. SecA or SecA(I304A/L306A) and wild type or PrlA4/SecYEG were used (n = 4–15). **x**, Mutant derivative. Mutated residues are indicated in capitals. **b**, The K_{cat} values (pmoles Pi per pmol SecA protomer per min) of the translocation ATPase activity of SecA at 37 °C divided by those of the corresponding membrane ATPase activity (Supplementary Fig. 3a) represent the folds of stimulation achieved by the various preprotein segments as indicated (n = 4–15).

to 2). Presumably, *prl* mutations mimic signal-peptide-induced triggering by stabilizing the same low energy conformational state of the translocase.

Prl mutants enabled us to examine the requirements of post-triggering 'secretion' reactions. PrlA4 mutant secreted more wild-type proPhoA *in vitro* than the wild-type translocase (Fig. 1c and d; compare lane 5 to 1)[19], indicating that 'triggering' is a rate-limiting step for secretion. PrlA4 (refs 10, 16 and 19) associated with PhoA, proPhoA(L8Q) and proPhoA(L14R) indistinguishably from the wild-type translocase (Fig. 1a), but secreted these preproteins 4–10 times more either *in vivo* or *in vitro* (Fig. 1b–d, compare lanes 2–4 to lanes 6–8). Nevertheless, secretion of mutant substrates, by either wild type or PrlA4 translocase, is still severely compromised compared to that of wild-type proPhoA (Fig. 1b–d; compare lanes 6–8 to lane 5)[9]. Clearly, *prl* mutations only bypass 'triggering'. For efficient protein 'secretion', even the PrlA4 translocase requires functional signal peptides.

These data show that 'secretion' requires the physical presence of signal peptides. To test this directly we used *in vitro* reconstitution. Wild-type synthetic peptide added *in trans* stimulated PhoA translocation into wild-type SecYEG IMVs (Fig. 3a, compare lane 7 to lane 5) to levels comparable to those seen with proPhoA (lane 3) and translocation ATPase (Fig. 2b, lane 8). Similar results were obtained with the PrlA4 translocase (Supplementary Fig. 5). Signal peptides that are defective for proPhoA secretion (Fig. 1b–d, lanes 2 and 3) are either severely (L8Q; Fig. 3a, lane 9) or completely (L14R; Fig. 3a, lane 10) compromised in driving PhoA secretion *in trans*. Signal-peptide-stimulated PhoA translocation requires physiological temperature (Fig. 3a, compare lane 8 to 7), ATP (Fig. 3a, compare lane 7 to 6) and the 'non-native' state of PhoA (Supplementary Fig. 6b), depends on signal peptide concentration (Supplementary Fig. 6a), proceeds with similar kinetics to those of proPhoA (Supplementary Fig. 6b) and is not affected by the order of ligand addition (Supplementary Fig. 5). Clearly, signal peptides are essential after triggering to drive mature domain secretion. For this role, their covalent linkage to mature chains is unnecessary.

To identify post-triggering steps required for secretion, complexes of [^{35}S]-PhoA bound to SecA–SecYEG IMVs were isolated (Fig. 3b). Excess of unlabelled PhoA readily replaces bound [^{35}S]-PhoA and prevents signal peptide-driven translocation seen in the absence of chase (Fig. 3b, compare lane 2 to 1). However, 2 min pre-incubation of bound [^{35}S]-PhoA with signal peptide before chase completely prevented exchange with unlabelled PhoA (Fig. 3b, compare lane 3 to 2). Presumably, signal peptides cause mature domains to become physically 'trapped' in the translocase. Trapping requires concomitant incubation with signal peptide and ATP at 37 °C (Fig. 3b, lane 3). Low temperature (lane 4), ATP-alone (lane 5), non-hydrolysable ATP (lanes 6 and 7) or a non-functional signal peptide (lane 8), all failed to drive the reaction. Identical results were obtained with the PrlA4 translocase (Fig. 3b, lanes 11–12). Trapped PhoA represents an early translocation intermediate because, first, trapping is reversed readily after brief chilling (lane 9) and second, all PhoA that is trapped at 2 min is protease-accessible (lane 10) and therefore still largely exposed to the cytoplasm.

The signal peptide binding groove on SecA[4] is essential for all of the sub-reactions dissected here. Inactivation of the signal peptide binding cleft reduces proPhoA affinity to that of PhoA (Fig. 1a) and severely compromises triggering (Fig. 2a, compare lane 11 to 3) as well as trapping (Fig. 3c, compare lane 4 to 2). Trapping remains defective (Fig. 3c, lane 6) even after use of a PrlA4 translocase to impose the triggered state artificially (Fig. 2a, lane 12).

The properties shown here are not only valid for proPhoA but are likely to be universal. Two other signal peptides, from proLamB[4] and proM13coat[20], drive triggering (Supplementary Fig. 7) and mature PhoA secretion (Fig. 3a, lanes 12 and 13). Four other mature domains bind to SecA with high affinity in the absence of signal peptides (Fig. 4a), whereas addition of the proPhoA signal peptide *in trans*

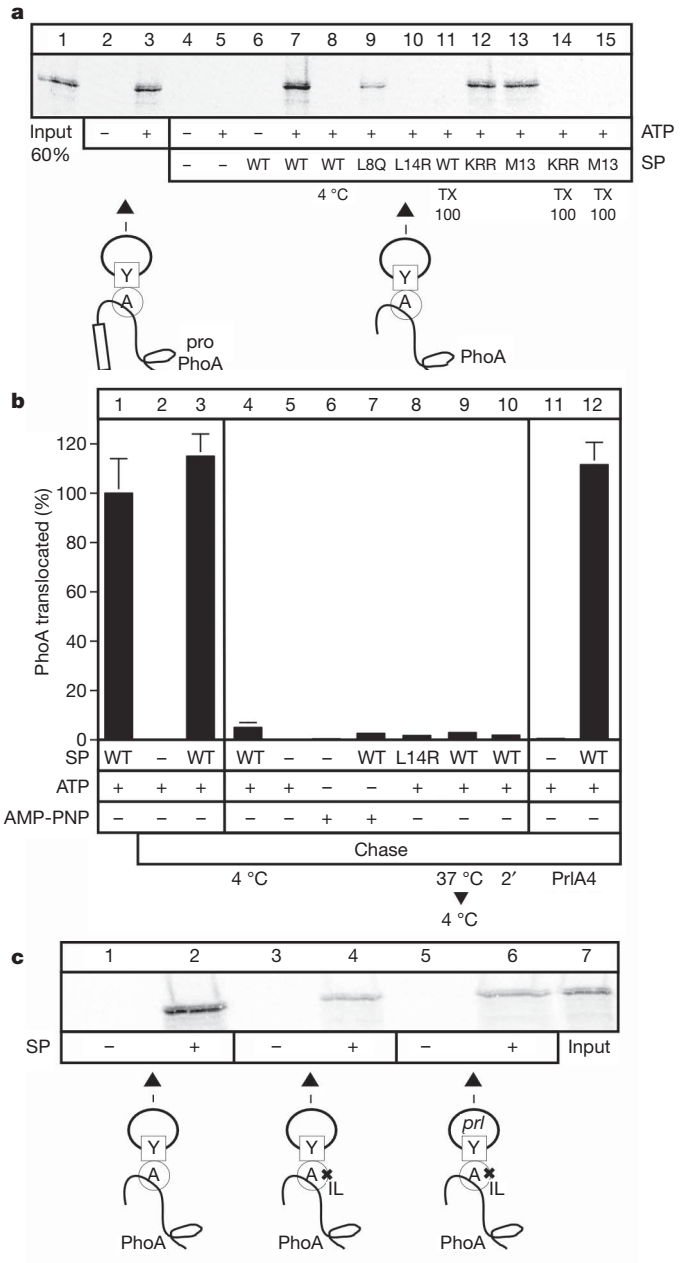


Figure 3 | Signal peptides added *in trans* promote PhoA translocation. **a**, [^{35}S]-PhoA translocation into wild-type SecYEG IMVs driven by proPhoA (WT, L8Q, L14R), M13 procoat or proLamB12 (KRR) signal peptides. Lanes 11, 14, 15 are identical to 7, 12, 13 except TX-100 was added before proteolysis. Lane 3, 100%; lane 7, 120% (± 16); lane 9, 7% (± 4); lane 12, 78% (± 10); lane 13, 71% (± 8); $n = 3$. **b**, Trapping reaction. Translocase was incubated with [^{35}S]-PhoA and then with nucleotides and/or with signal peptides. Where previously omitted, ATP and/or signal peptide were added. Samples (except lane 1) were chased with non-radiolabelled PhoA at 37 °C (except lane 4). Lane 9: after 2 min the reaction was chilled (4 °C) before translocation resumed. Lane 1, 100% ($n = 3$). AMP-PNP, Adenylyl-imidodiphosphate. **c**, proPhoA signal-peptide-driven [^{35}S]-PhoA translocation (as in **a**). Lane 2, 100%; lane 4, 5% (± 1.1); lane 6, 12% (± 2.4); $n = 3$.

drives their secretion (Fig. 4b) as efficiently as their own signal peptides (Fig. 4c).

Preprotein signal peptides and mature domains have several distinct roles in secretion (Fig. 4d)[15,20,21]. Signal peptides and, in some cases, SecB binding stabilize 'non-native' states and preproteins are then targeted to the translocase[5,6,8,22]. An additional targeting route, universal in bacteria, was shown here (Fig. 4d, targeting). This involves direct recognition of 'non-native' mature domains by cytoplasmic or SecYEG-bound SecA. SecA is ubiquitous in Bacteria and

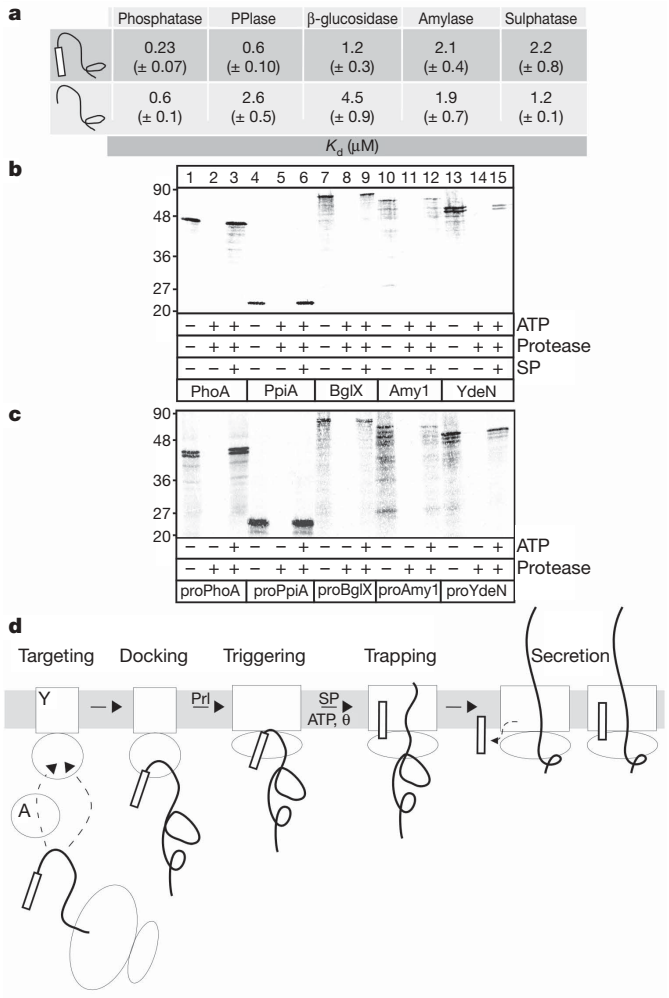


Figure 4 | Generality and working model of bacterial secretory protein translocation. a, Equilibrium dissociation constants (K_d) of precursor and mature forms of the indicated secretory *Escherichia coli* proteins for the translocase. **b**, **c**, *In vitro* translocation reactions containing [^{35}S]-labelled mature forms and synthetic proPhoA signal peptide (**b**) or [^{35}S]-labelled precursor forms of the indicated secretory proteins (**c**, as in Fig. 3a). **d**, Model of post-translational bacterial protein secretion (see text for details). A nascent secretory chain (thick line) carrying a signal peptide (rectangle) is shown to exit the ribosome. A, SecA; Y, SecY. Elongated shapes depict the triggered conformational state.

shuttles between cytoplasm and membrane[5]. Mature domain targeting signals could be degenerate sequences that become buried in 'native' structures, reminiscent of chaperone recognition[23,24].

Mature domains are main contributors to the docking of several preproteins on SecA (Fig. 4a, d, docking). This prominence was previously unsuspected. In some cases, signal peptides slightly enhance preprotein binding (Fig. 4a, see phosphatase, β-glucosidase, peptidyl-prolyl *cis-trans* isomerase (PPIase)), in others they have no contribution (see amylase and sulphatase). Signal peptides with higher affinities[14,20] or those attached to short mature domains (for example, proPhoA(1–62); Fig. 1a) could contribute more to preprotein docking. Mature domains (Fig. 1a)[15,25] and signal peptides[4] dock at non-overlapping SecA sites (Supplementary Fig. 2). These must be proximal because proPhoA, in which signal peptide and mature domain are covalently connected, binds two- to threefold more tightly than PhoA (Fig. 1a). Being significantly larger, mature domains might facilitate positioning of signal peptides over their SecA binding cleft[4].

Tight signal peptide binding to SecA promotes triggering of the translocase holoenzyme (Fig. 4d, triggering) possibly by priming the protein-conducting channels[11,26,27] for opening. Next, it drives trapping (Fig. 4d, trapping) of the first amino-terminal segment of mature PhoA in the translocase so that mature domains become irreversibly engaged in the channel. Trapping is tightly coupled to subsequent complete secretion (Fig. 4d, secretion) through several rounds of ATP hydrolysis and engagement of succeeding mature domain segments. Signal peptides could come off after trapping (Fig. 4d, secretion, left) or may remain bound throughout secretion (Fig. 4d, secretion, right). Following triggering, signal peptides are expected to form additional intimate interactions with the SecYEG channel[18,28].

This cascade of events imposes several checkpoints that ensure efficient sorting of secretory proteins from cytoplasmic residents. Cytoplasmic proteins fold rapidly and will not be recognized by SecA. Without a signal peptide, the occasional illicit cytoplasmic binder cannot trigger the translocase and as a result of this proofreading-like function it will be rejected[17,18].

METHODS SUMMARY

Strains and reagents. Bacterial strains expressing proPhoA and derivatives have been described previously[14,19]. SecA and urea-treated IMVs were prepared as described[29]. SecY amounts were quantified using western blots with anti-SecY immunostaining. All genetic constructs and antibodies are described in the Supplementary Methods. Preproteins were purified by Ni^{+2} affinity chromatography under denaturing conditions in buffer C (50 mM Tris-HCl pH 8.0; 50 mM KCl; 6 M urea; 10% glycerol v/v) and were stored in buffer D (50 mM Tris-HCl pH 8.0; 50 mM KCl; 6 M urea; 1 mM EDTA; 10% glycerol v/v). Alkaline phosphatase units were determined *in vivo* using *p*-nitrophenol phosphate (Sigma) as described[10,13] and were converted to secreted protein mass using a standard curve obtained by determining the units of increasing amounts of purified native PhoA. Signal peptides (chemically synthesized; GenScript) were stored at 15 mM in 100% dimethylsulphoxide at 4 °C.

Biochemical and biophysical assays. [^{35}S]-labelled proPhoA and derivatives were prepared by *in vitro* transcription/translation (Promega) and bound to inverted inner membrane vesicles as described[29]. Binding of proPhoA and derivatives to SecA in solution was determined by isothermal titration calorimetry (VP-ITC, MicroCal) at 8 °C as described[4]. ProPhoA and derivatives were kept in the ITC measuring cell (80 μM; 20 mM Tris-HCl pH 8.0, 20 mM KCl, supplemented with 2 mM Tris (2-carboxyethyl) phosphine (TCEP) to maintain a non-native state), while SecA (1 mM) was added in 20 μl injection steps. Thermal ATPase assays were performed as described[14] in buffer B (50 mM Tris-HCl pH 8.0; 50 mM KCl; 5 mM $MgCl_2$) supplemented with 0.4 μM SecA; 0.5 mg ml^{-1} BSA; 1 mM ATP and 1.5 mM dithiothreitol (unless otherwise specified). For membrane ATPase, IMVs (0.4 μM SecY) were added. For translocation ATPase, proPhoA or derivatives were further added at indicated amounts. Activation energies were derived from Arrhenius transformations (Supplementary Fig. 4c).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 8 July; accepted 6 October 2009.

1. Blobel, G. & Dobberstein, B. Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. *J. Cell Biol.* **67**, 835–851 (1975).
2. Rapoport, T. A. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature* **450**, 663–669 (2007).
3. Papanikou, E., Karamanou, S. & Economou, A. Bacterial protein secretion through the translocase nanomachine. *Nature Rev. Microbiol.* **5**, 839–851 (2007).
4. Gelis, I. *et al.* Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR. *Cell* **131**, 756–769 (2007).
5. Hartl, F. U., Lecker, S., Schiebel, E., Hendrick, J. P. & Wickner, W. The binding cascade of SecB to SecA to SecY/E mediates preprotein targeting to the *E.coli* plasma membrane. *Cell* **63**, 269–279 (1990).
6. Zhou, J. & Xu, Z. The structural view of bacterial translocation-specific chaperone SecB: implications for function. *Mol. Microbiol.* **58**, 349–357 (2005).
7. Schiebel, E., Driessen, A. J., Hartl, F. U. & Wickner, W. $\Delta\mu_H+$ and ATP function at different steps of the catalytic cycle of preprotein translocase. *Cell* **64**, 927–939 (1991).
8. Shimizu, H., Nishiyama, K. & Tokuda, H. Expression of *gpsA* encoding biosynthetic *sn*-glycerol 3-phosphate dehydrogenase suppresses both the LB^- phenotype of a *secB* null mutant and the cold-sensitive phenotype of a *secG* null mutant. *Mol. Microbiol.* **26**, 1013–1021 (1997).
9. Prinz, W. A., Spiess, C., Ehrmann, M., Schierle, C. & Beckwith, J. Targeting of signal sequenceless proteins for export in *Escherichia coli* with altered protein translocase. *EMBO J.* **15**, 5209–5217 (1996).

10. Derman, A. I., Puziss, J. W., Bassford, P. J. Jr & Beckwith, J. A signal sequence is not required for protein export in *prlA* mutants of *Escherichia coli*. *EMBO J.* **12,** 879–888 (1993).
11. Li, W. *et al.* The plug domain of the SecY protein stabilizes the closed state of the translocation channel and maintains a membrane seal. *Mol. Cell* **26,** 511–521 (2007).
12. Cooper, D. B. *et al.* SecA, the motor of the secretion machine, binds diverse partners on one interactive surface. *J. Mol. Biol.* **382,** 74–87 (2008).
13. Akiyama, Y. & Ito, K. Folding and assembly of bacterial alkaline phosphatase *in vitro* and *in vivo*. *J. Biol. Chem.* **268,** 8146–8150 (1993).
14. Karamanou, S. *et al.* Preprotein-controlled catalysis in the helicase motor of SecA. *EMBO J.* **26,** 2904–2914 (2007).
15. Lill, R., Dowhan, W. & Wickner, W. The ATPase activity of SecA is regulated by acidic phospholipids, SecY, and the leader and mature domains of precursor proteins. *Cell* **60,** 271–280 (1990).
16. Smith, M. A., Clemons, W. M. Jr, DeMars, C. J. & Flower, A. M. Modeling the effects of *prl* mutations on the *Escherichia coli* SecY complex. *J. Bacteriol.* **187,** 6454–6465 (2005).
17. Osborne, R. S. & Silhavy, T. J. PrlA suppressor mutations cluster in regions corresponding to three distinct topological domains. *EMBO J.* **12,** 3391–3398 (1993).
18. Flower, A. M., Doebele, R. C. & Silhavy, T. J. PrlA and PrlG suppressors reduce the requirement for signal sequence recognition. *J. Bacteriol.* **176,** 5607–5614 (1994).
19. van der Wolk, J. P. *et al.* PrlA4 prevents the rejection of signal sequence defective preproteins by stabilizing the SecA–SecY interaction during the initiation of translocation. *EMBO J.* **17,** 3631–3639 (1998).
20. Papanikou, E. *et al.* Identification of the preprotein binding domain of SecA. *J. Biol. Chem.* **280,** 43209–43217 (2005).
21. Carlsson, F. *et al.* Signal sequence directs localized secretion of bacterial surface proteins. *Nature* **442,** 943–946 (2006).
22. Park, S., Liu, G., Topping, T. B., Cover, W. H. & Randall, L. L. Modulation of folding pathways of exported proteins by the leader sequence. *Science* **239,** 1033–1035 (1988).
23. Stan, G., Brooks, B. R., Lorimer, G. H. & Thirumalai, D. Residues in substrate proteins that interact with GroEL in the capture process are buried in the native state. *Proc. Natl Acad. Sci. USA* **103,** 4433–4438 (2006).
24. Rudiger, S., Germeroth, L., Schneider-Mergener, J. & Bukau, B. Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. *EMBO J.* **16,** 1501–1507 (1997).
25. Erlandson, K. J. *et al.* A role for the two-helix finger of the SecA ATPase in protein translocation. *Nature* **455,** 984–987 (2008).
26. Simon, S. M. & Blobel, G. Signal peptides open protein-conducting channels in *E.coli*. *Cell* **69,** 677–684 (1992).
27. Saparov, S. M. *et al.* Determining the conductance of the SecY protein translocation channel for small molecules. *Mol. Cell* **26,** 501–509 (2007).
28. Osborne, A. R. & Rapoport, T. A. Protein translocation is mediated by oligomers of the SecY complex with one SecY copy forming the channel. *Cell* **129,** 97–110 (2007).
29. Vrontou, E., Karamanou, S., Baud, C., Sianidis, G. & Economou, A. Global co-ordination of protein translocation by the SecA IRA1 switch. *J. Biol. Chem.* **279,** 22490–22497 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to G. Dimitrakopoulos for help with MATLAB; M. Koukaki for constructs and materials; D. Boyd, S. Schulman and V. Koronakis for gifts of strains, plasmids and protocols; A. Kohen for advice on Arrhenius plots; A. Kuhn and K. Tokatlidis for comments. The research leading to these results has received funding from the European Community's Sixth Framework Programme (FP6/2002–2007) under grant agreement LSHC-CT-2006-037834/Streptomics (to A.E.), the Greek General Secretariat of Research and the European Regional Development Fund (PENED03ED623; to A.E.), the US National Institutes of Health grant GM73854 (to C.G.K.) and a Scientist Development Grant by the American Heart Association (to C.G.K.). G.G. is an Onassis Foundation predoctoral fellow.

Author Contributions G.G. cloned genes, performed *in vivo* and *in vitro* secretion experiments, phosphatase assays, membrane binding studies, Arrhenius conversions and developed the *in trans* reconstitution assay. G.G. and S.K. purified proteins, performed ATPase experiments, analysed data, provided experimental ideas and contributed in writing the paper. S.K. developed thermal-dependence ATPase assay, contributed in assay development, performed preliminary ITC experiments and edited the paper. I.G. purified proteins, performed and analysed ITC experiments. C.G.K. designed, guided and analysed ITC experiments, contributed in experimental ideas and controls and in writing and editing the paper. A.E. conceived, designed and guided experiments, analysed data and wrote the paper. All authors read and commented on the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.E. (aeconomo@imbb.forth.gr).

METHODS

Determination of equilibrium dissociation constants (K_d) of proPhoA variants for SecYEG-bound SecA. Non-radiolabelled proPhoA variants were serially diluted in buffer E (50 mM Tris-HCl pH 8.0; 50 mM KCl; 6 M urea; 1 mM EDTA; 1 mM DTT; 5% glycerol v/v). An aliquot (0.7 μl) from each dilution was added to the final reaction (20 μl) to achieve the desired protein concentration. Non-radiolabelled proPhoA variants were added in a concentration range of 1–30,000 μM (depending on the K_d of each variant for the SecYEG-bound SecA). IMVs containing overexpressed SecYEG (2 mg ml^{-1} total membrane protein) were diluted 25 times in buffer B (50 mM Tris-HCl pH:8.0; 50 mM KCl, 5 mM $MgCl_2$), and 5 μl of these were mixed with 4.3 μl of SecA (0.2 μg μl^{-1}). The SecA–SecYEG complex (9.3 μl) was incubated on ice for 10 min and then added in the reactions. An aliquot (5 μl) of the reaction mix (prepared by mixing 0.8 ml of BSA 100 mg ml^{-1}; 8 ml of 10× buffer B and distilled water up to 20 ml) and 5 μl of [^{35}S]-proPhoA or its variants' dilution were also added to the reactions. Reactions were then incubated on ice for 20 min, overlaid on an equal volume of BSA/sucrose cushion (prepared by mixing 1.37 g sucrose with 5 ml of the reaction mix and made up to 20 ml with distilled water) and ultracentrifuged (320,000*g*; 30 min; 4 °C). The membrane bound material in the pellet was isolated and resuspended in buffer B (50 mM Tris-HCl pH:8.0; 50 mM KCl, 5 mM $MgCl_2$) and then immobilized by spotting the resuspended pellets on a nitrocellulose membrane using a vacuum manifold (Bio-Rad). Data were analysed by nonlinear regression fitting for one binding site using Prism (Graph Pad) as described[29]. IMVs were prepared from cells that overexpress either a wild-type *secY/secE/secG* operon[30] or the *prlA4* (that is, *secY(I408N/F286Y)/secE/secG*, see ref. 16) operon. SecA and SecA(I304A/L306A) (mutated in the signal peptide binding groove[4]), were prepared as described. After synthesis of [^{35}S]-proPhoA variants, buffer exchange was accomplished using G-50 resin equilibrated with buffer E (50 mM Tris-HCl pH 8.0; 50 mM KCl; 6 M urea; 1 mM EDTA; 1 mM DTT; 5% glycerol v/v) and an ~eightfold dilution was made in buffer B supplemented with 1 mM DTT before adding them in the reactions. Non-radiolabelled proPhoA variants were treated with 2 mM DTT for 8 h to maintain them in a 'non-native', translocation-competent state.

Determination of equilibrium dissociation constants (K_d) of proPhoA variants and chemically synthesized proPhoA signal peptide for free SecA. K_d determination of proPhoA variants and chemically synthesized proPhoA signal peptide for free SecA was determined by ITC as described[4] and explained in the legend to Supplementary Fig. 3.

***In vivo* translocation of proPhoA and its derivatives.** The wild type *secY/secE/secG* operon[30] or the *prlA4* (that is, *secY(I408N/F286Y)/secE/secG*, see ref. 16) operon was cloned in pET610 plasmid which is under the control of a *trc* promoter[30]. Genes encoding proPhoA and its derivatives were cloned in the compatible pBAD33 plasmid, under the control of the arabinose promoter[11]. The two plasmids were co-transformed in JM109 cells. Cells were grown at 37 °C until $D_{595} = 0.2$, SecYEG synthesis was induced by addition of 0.2 mM IPTG (isopropylthiogalactoside), while synthesis of proPhoA and derivatives was based on the read-through of the arabinose promoter[11] in the absence of arabinose. After IPTG induction (20 min), cells were pelleted by centrifugation (3,834 *g*; 4 min) and resuspended in 1 M Tris pH:8.0 followed by addition of *p*-nitrophenyl phosphate (15 mM). The reaction was incubated at 37 °C for an appropriate time until a strong yellow colour was observed and then stopped with a 10% (v/v) of a solution obtained by mixing 1 volume of 0.5 M EDTA pH 8.0 and 4 volumes of 2.5 M K_2HPO_4. After addition of the non-ionic detergent TX-100 (1% v/v) and centrifugation (17,000 *g*; 4 min) in order to remove cell debris, the absorbance of *p*-nitrophenol was determined at 420 nm. Units of alkaline phosphatase were calculated as described[10] and converted to mass of secreted protein by using a standard curve with purified PhoA. Secretion of the proPhoA derivatives mediated by chromosomal SecYEG was measured under the same conditions using an empty pET610 plasmid and these values were subtracted.

***In vitro* translocation of proPhoA and its derivatives.** Reactions were performed in 100 μl buffer B; 0.5 mg ml^{-1} BSA, 2.5 mM ATP, 1 mM DTT by addition of SecA (0.4 μM), wild-type or PrlA4-SecYEG IMVs (1.0 μM SecY) and proPhoA or its derivatives (8.5 μM). Reactions were incubated at 37 °C for 12 min and translocation into the lumen of the IMVs was terminated by addition of proteinase K (1 mg ml^{-1}, 20 min, 4 °C). Proteins were precipitated with trichloroacetic acid (TCA; 15% w/v), analysed by SDS–PAGE (13% acrylamide) and immunostained with anti-PhoA antibody. In all cases quantification is carried out compared to the stain of a fraction of the protease-untreated [^{35}S]-PhoA material.

Calculation of activation energies (E_a). The activation energies (E_a) of the translocase were derived from Arrhenius plots using measured K_{cat} values (pmoles Pi per pmol SecA protomer per min) of basal, membrane and translocation ATPase activities of SecA, as a function of temperature[14] (as described in Supplementary Fig. 4c). The *y* axis in the Arrhenius plots represents the natural logarithm of the K_{cat} values and the *x* axis the inversed temperature values (1/T) expressed in Kelvin. The activation energies (E_a) of the translocase under different regimes were calculated (in kJ mole^{-1}) using the slopes of the linear parts of the curves. SecA or SecA(I304A/L306A) mutants were used at 0.4 μM; wild-type or PrlA4/SecYEG at 0.4 μM SecY; proPhoA derivatives at 8.5 μM; synthetic signal peptides at 15 μM.

***In vitro* reconstitution of [^{35}S]-PhoA translocation into wild-type SecYEG IMVs by *in trans* addition of synthetic signal peptides.** Reactions were performed in 100 μl buffer B; 0.5 mg ml^{-1} BSA, 2.5 mM ATP, 1 mM DTT by addition of SecA (0.4 μM), SecYEG IMVs (1.0 μM SecY), freshly prepared [^{35}S]-PhoA or [^{35}S]-proPhoA (~300 fmoles) and synthetic signal peptides (50 μM). Reactions were incubated at 37 °C for 12 min and translocation into the lumen of the IMVs was terminated by proteinase K addition (1 mg ml^{-1}; 20 min; 4 °C). Proteins were precipitated with trichloroacetic acid (TCA; 15% w/v) and analysed by SDS–PAGE (13% acrylamide). Molecular masses (kDa) were derived from five marker proteins (from top to bottom: β-galactosidase, bovine serum albumin, ovalbumin, carbonic anhydrase, β-lactoglobulin, lysozyme). The gel was incubated with 1 M sodium salicylate (1 h) and then visualized by phosphorimaging. Signal peptides (chemically synthesized; GenScript) were stored at 15 mM in 100% dimethylsulphoxide at 4 °C, diluted at 1 mM in 10 mM Tris-HCl pH 7.0 before being added to the reactions. In all cases quantification is carried out compared to the stain of a fraction of the protease-untreated [^{35}S]-PhoA material taken as 100%. In Fig. 4c. samples in lanes 8 and 9, 11 and 12, 14 and 15 were loaded with three times more material than those of lanes 7, 10 and 13, respectively, and were quantified as follows: in b, lane 3, 180% (± 20); lane 6, 61% (± 10); lane 9, 16% (± 8); lane 12, 20% (± 6); lane 15, 4% (± 2). In c, lane 3, 165% (± 35); lane 6, 110% (± 12); lane 9, 22% (± 4); lane 12, 17% (± 8); lane 15, 13% (± 4).

Trapping of the polypeptide chain in the translocase holoenzyme. The translocase holoenzyme, assembled on SecYEG IMVs (1.0 μM SecY) by addition of 0.4 μM SecA in buffer B, was incubated on ice for 10 min with freshly prepared [^{35}S]-PhoA (~600 fmoles), overlaid on an equal volume of BSA/sucrose cushion (prepared as previously described) and ultracentrifuged (320,000*g*; 30 min; 4 °C). The SecYEG bound [^{35}S]-PhoA present in the pellet was isolated and resuspended in buffer B and then incubated for 2 min at 37 °C with nucleotides (ATP, AMP-PNP (adenylyl-imidodiphosphate); 1 mM) and/or synthetic proPhoA signal peptides (wild type or L14R; 50 μM). Where previously omitted, reactions were supplemented with ATP (1.5 mM) and/or proPhoA signal peptide (50 μM) to initiate translocation into the lumen of IMVs by transfer at 37 °C as previously described. At the same time all reactions were chased with excess of non-radiolabelled PhoA (1.5 μM). Translocation into the lumen of the IMVs was terminated by proteinase K addition (1 mg ml^{-1}; 20 min; 4 °C). Proteins were precipitated with trichloroacetic acid (TCA; 15% w/v), analysed by SDS–PAGE (13% acrylamide). The gel was incubated with sodium salicilate (1 h, 1 M) and then visualized by phosphorimaging.

30. van der Does, C. *et al.* SecA is an intrinsic subunit of the *Escherichia coli* preprotein translocase and exposes its carboxyl terminus to the periplasm. *Mol. Microbiol.* **22**, 619–629 (1996).

LETTERS

Dynamic activation of an allosteric regulatory protein

Shiou-Ru Tzeng[1,2] & Charalampos G. Kalodimos[1,2]

Allosteric regulation is used as a very efficient mechanism to control protein activity in most biological processes, including signal transduction, metabolism, catalysis and gene regulation[1–6]. Allosteric proteins can exist in several conformational states with distinct binding or enzymatic activity. Effectors are considered to function in a purely structural manner by selectively stabilizing a specific conformational state, thereby regulating protein activity. Here we show that allosteric proteins can be regulated predominantly by changes in their structural dynamics. We have used NMR spectroscopy and isothermal titration calorimetry to characterize cyclic AMP (cAMP) binding to the catabolite activator protein (CAP), a transcriptional activator that has been a prototype for understanding effector-mediated allosteric control of protein activity[7]. cAMP switches CAP from the 'off' state (inactive), which binds DNA weakly and non-specifically, to the 'on' state (active), which binds DNA strongly and specifically. In contrast, cAMP binding to a single CAP mutant, CAP-S62F, fails to elicit the active conformation; yet, cAMP binding to CAP-S62F strongly activates the protein for DNA binding. NMR and thermodynamic analyses show that despite the fact that CAP-S62F-cAMP$_2$ adopts the inactive conformation, its strong binding to DNA is driven by a large conformational entropy originating in enhanced protein motions induced by DNA binding. The results provide strong evidence that changes in protein motions may activate allosteric proteins that are otherwise structurally inactive.

The cAMP-mediated activation of CAP for DNA binding[8–10] is a characteristic example of the widely held view that allosteric regulation is predominantly structural in origin[6,11]. The DNA-binding domain (DBD; residues 139–209) in cAMP-free CAP (apo-CAP) adopts an orientation that is incompatible with DNA binding[9]. cAMP binding to the cAMP-binding domain (CBD; residues 1–135) of CAP (Supplementary Fig. 1) elicits allosterically a pronounced conformational change to DBD, which undergoes a ~60° rigid-body rotation (Fig. 1a and Supplementary Fig. 2). In this orientation, the recognition helices (F-helices) are optimally poised to interact with the major groove of the DNA[8,10] and the affinity of CAP for DNA is enhanced by several orders of magnitude[12]. Therefore, the structural basis for cAMP-induced CAP activation consists of a marked alteration of the relative orientation of DBD[9].

The NMR spectra of CAP in all of its three functional states (unliganded, cAMP$_2$-bound, and cAMP$_2$-DNA-bound) are excellent (Supplementary Fig. 3). The pronounced effect elicited by cAMP binding to wild-type (WT)-CAP structure[9,13] is reflected in the widespread changes in the chemical shift ($\Delta\omega$; Fig. 1b and Supplementary Figs 4 and 5). DNA binding to CAP-cAMP$_2$ has a much less pronounced chemical shift effect than cAMP binding to CAP and is mainly confined to the DBD (Supplementary Fig. 4a), in agreement with structural data showing that CAP-cAMP$_2$ undergoes very little conformational change when interacting with the DNA[8,10] (Fig. 1a). Thus, the distinct chemical shifts of DBD in apo-CAP, wherein DBD is in the inactive conformation, and CAP-cAMP$_2$, wherein DBD is in the fully active state, can serve as a proxy for the active and inactive DNA-binding conformational states (Supplementary Fig. 6).

The CAP-S62F mutant (Supplementary Fig. 7) was originally isolated *in vivo* along with a number of constitutively active mutants (termed CAP* mutants)[14]. However, in contrast to other CAP* mutants, CAP-S62F was reported not to enhance CAP affinity for DNA in the absence of cAMP[12]. Addition of cAMP, however, was reported to render CAP-S62F as active as WT-CAP[12,14]. NMR analysis clearly demonstrates that the DBD in apo-CAP-S62F is in the inactive, DNA-binding incompetent conformation (Fig. 1c and Supplementary Figs 4b and 6). The NMR data show that cAMP binding to CAP-S62F causes extensive chemical shift changes to CBD (Fig. 1c and Supplementary Figs 4b and 6), similar, albeit less pronounced, to the effect of cAMP binding to CBD of WT-CAP (Fig. 1b and Supplementary Fig. 4a).

Notably, cAMP binding to CAP-S62F seems to have no effect on the conformation of the DBD, which seems to remain in the inactive state (Fig. 1c and Supplementary Figs 4b and 6). In fact, the conformation of the DBD in the CAP-S62F-cAMP$_2$ complex is very similar, if not identical, to the one that DBD adopts in the unliganded protein (either WT-CAP or CAP-S62F). Apparently, the S62F substitution decouples the long-range structural communication between CBD and DBD and as a result the relative orientation of DBD is not allosterically affected by cAMP binding to CAP-S62F (Supplementary Fig. 7). Therefore, whereas cAMP binding to WT-CAP induces the active, DNA-binding competent conformation, cAMP binding to CAP-S62F fails to induce the active DBD conformation.

The observation that DBD in CAP-S62F-cAMP$_2$ exists in the inactive conformation, and hence would be expected not to bind to DNA, is clearly at odds with earlier observations that CAP-S62F-cAMP$_2$ does enhance transcription *in vivo* and binds tightly to DNA *in vitro*[12,14]. Indeed, NMR analysis shows that both WT-CAP-cAMP$_2$ and CAP-S62F-cAMP$_2$ form strong complexes with DNA (Supplementary Fig. 8). Chemical shift differences are negligible, indicating that the two DNA complexes have very similar structure, with the exception of the region surrounding the single point substitution (S62F) in the cAMP-binding pocket where some notable differences are observed (Supplementary Figs 6 and 8).

To determine the thermodynamic basis for WT-CAP-cAMP$_2$ and CAP-S62F-cAMP$_2$ binding to DNA, we used high-sensitivity isothermal titration calorimetry (ITC) to measure the association free energy (ΔG), and its enthalpic (ΔH) and entropic (ΔS) components. Both WT-CAP-cAMP$_2$ and CAP-S62F-cAMP$_2$ bind to DNA with very similar affinity (K_d ~0.4 μM compared with 0.6 μM, respectively; Fig. 2a). Notably, although the two proteins bind to DNA equally strongly, their thermodynamic basis for the interaction with DNA is fundamentally different. At physiological temperatures, WT-CAP-cAMP$_2$ binding to DNA is enthalpically favoured ($\Delta H = -23.2$ kcal mol^{-1}), and entropically opposed ($-T\Delta S = 14.3$ kcal mol^{-1}), whereas binding of CAP-S62F-cAMP$_2$ to DNA is driven entirely by a large increase in entropy ($-T\Delta S = -13.2$ kcal mol^{-1}) (Fig. 2a). Thus, the difference in entropy, $\Delta(-T\Delta S)$, for DNA binding to WT-CAP-cAMP$_2$ and CAP-S62F-cAMP$_2$ is enormous, amounting to 27.5 kcal mol^{-1}. It is of particular interest that the WT-CAP-cAMP$_2$ and CAP-S62F-cAMP$_2$ complexes

[1]Department of Chemistry & Chemical Biology, [2]Department of Biomedical Engineering, Rutgers University, Piscataway, New Jersey 08854, USA.

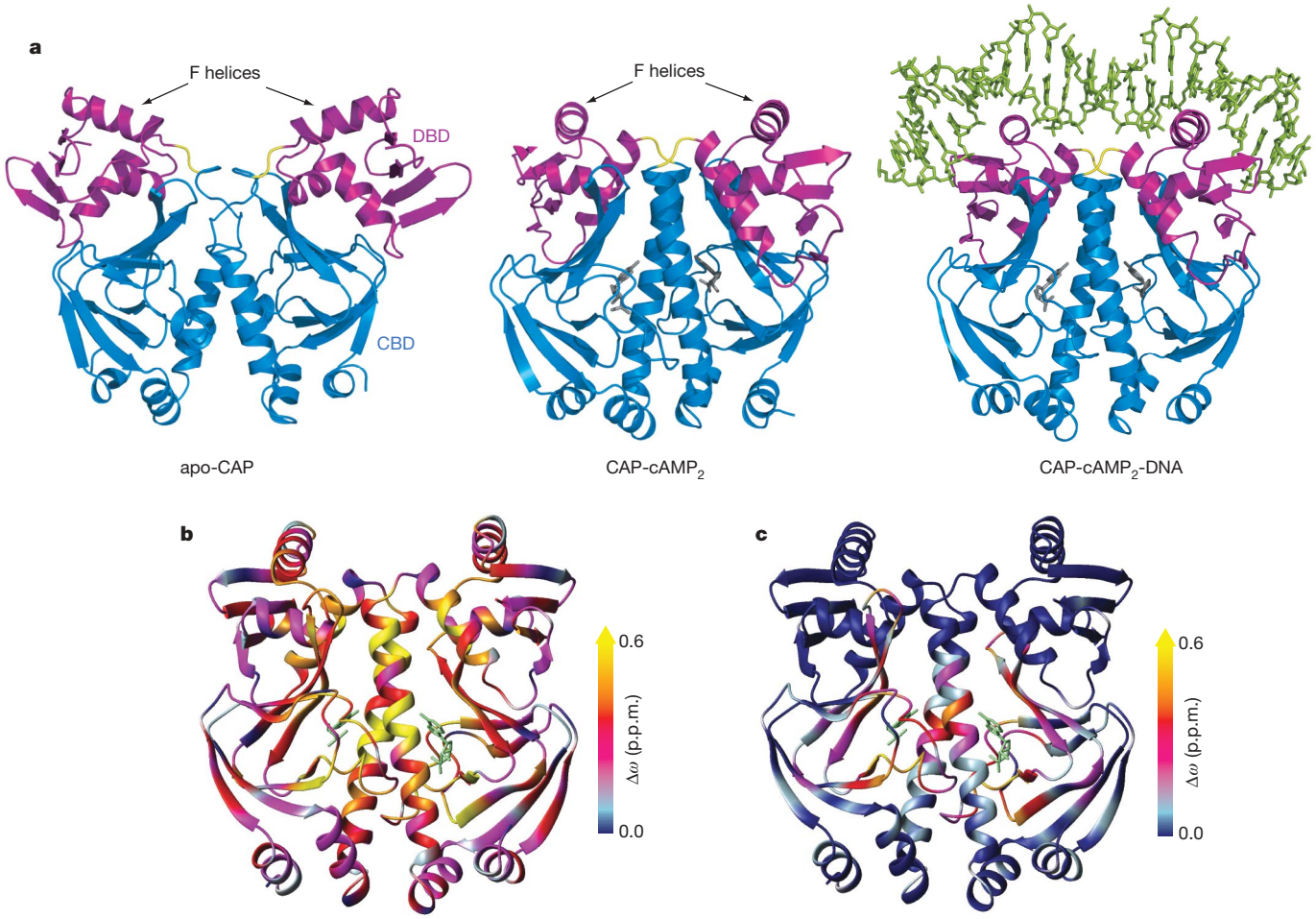


Figure 1 | Conformational states of CAP and effect of cAMP binding assessed by NMR. **a**, Structures of CAP in three ligation states: apo[9], $cAMP_2$-bound[10], and $cAMP_2$-DNA-bound[8]. The CBD, DBD and hinge region are coloured blue, magenta and yellow, respectively. cAMP and DNA are displayed as grey and green sticks, respectively. **b**, **c**, Effect of cAMP binding on the structure of WT-CAP (**b**) and CAP-S62F (**c**) as assessed by chemical shift mapping (Supplementary Fig. S4). Chemical shift difference ($\Delta\omega$; p.p.m.) values are mapped by continuous-scale colour onto the WT-CAP-$cAMP_2$ structure.

use distinct thermodynamic strategies to interact strongly and specifically with DNA.

To better understand the mechanism by which CAP-S62F-$cAMP_2$ manages to bind strongly to DNA while adopting the DNA-binding inactive conformation, we performed a series of relaxation dispersion experiments (Fig. 3a). These experiments have the capacity to detect and characterize low-populated conformations[15,16]. The results show that on binding of cAMP to CAP-S62F, DBD resonances become broader, indicating the presence of exchange between conformations on the micro-to-millisecond (μs–ms) time scale. Data fitting (see Methods) is indicative of a two-site exchange process, with the population of the excited state being ~2% (Fig. 3a). The additional line broadening of NMR signals (R_{ex}; Fig. 3c) caused by conformational exchange between the ground (A) and an excited state (B) depends on the relative populations of the exchanging species (p_A and p_B) and the chemical shift difference between the exchanging species ($\Delta\omega$)[15,16]. The absolute ^{15}N $\Delta\omega$ values of DBD residues measured between the apo-CAP and WT-CAP-$cAMP_2$ (Figs 1b and 3b) clearly correlate with the $\Delta\omega$ values between the major and the minor conformations of CAP-S62F-cAMP2 determined by relaxation dispersion measurements ($\Delta\omega_{disp}$; Fig. 3d). Thus, the data provide strong evidence that the excited state that DBD transiently populates in CAP-S62F-$cAMP_2$ closely resembles the active, DNA-binding compatible conformation. Because the affinity of the active DBD conformation for DNA (for example, in CAP-$cAMP_2$) is many orders of magnitude higher than that of the inactive DBD conformation (for example in apo-CAP), DNA will preferentially bind to the active DBD conformation of CAP-S62F-$cAMP_2$, despite being so poorly populated. Thus, the data indicate that DNA binding to CAP-S62F-$cAMP_2$ proceeds with a population-shift mechanism[17].

Despite adopting predominantly the inactive conformation and only very poorly the active one (~2%), CAP-S62F-$cAMP_2$ binds to DNA as tightly as WT-CAP-$cAMP_2$, driven by a large favourable binding entropy change, as measured experimentally by calorimetry (Fig. 2a). The amount of surface that becomes buried on binding of DNA to WT-CAP-$cAMP_2$ and CAP-S62F-$cAMP_2$ is very similar, indicating that the hydrophobic effect is not the source of the large entropy difference measured for the formation of the two DNA complexes. To understand the origin of this large favourable change in entropy, we sought to determine the role of dynamics in the binding process. To assess the contribution of protein motions to the conformational entropy of the system[18,19], we measured changes in N-H bond order parameters for DNA binding to WT-CAP-$cAMP_2$ and CAP-S62F-$cAMP_2$ (Supplementary Figs 9–13). The order parameter, S^2, is a measure of the amplitude of internal motions on the ps–ns timescale and may vary from $S^2 = 1$, for a bond vector having no internal motion, to $S^2 = 0$, for a bond vector rapidly sampling multiple orientations[20].

DNA binding to WT-CAP-$cAMP_2$ results in widespread increase in S^2, indicating a global rigidification of the protein (Fig. 2b and Supplementary Fig. 13c). Notably, DNA binding to CAP-S62F-$cAMP_2$ causes a large number of residues to increase their motions as evidenced by the corresponding decrease in their S^2 values (Fig. 2b and Supplementary Fig. 13c). It is of interest to note that changes in

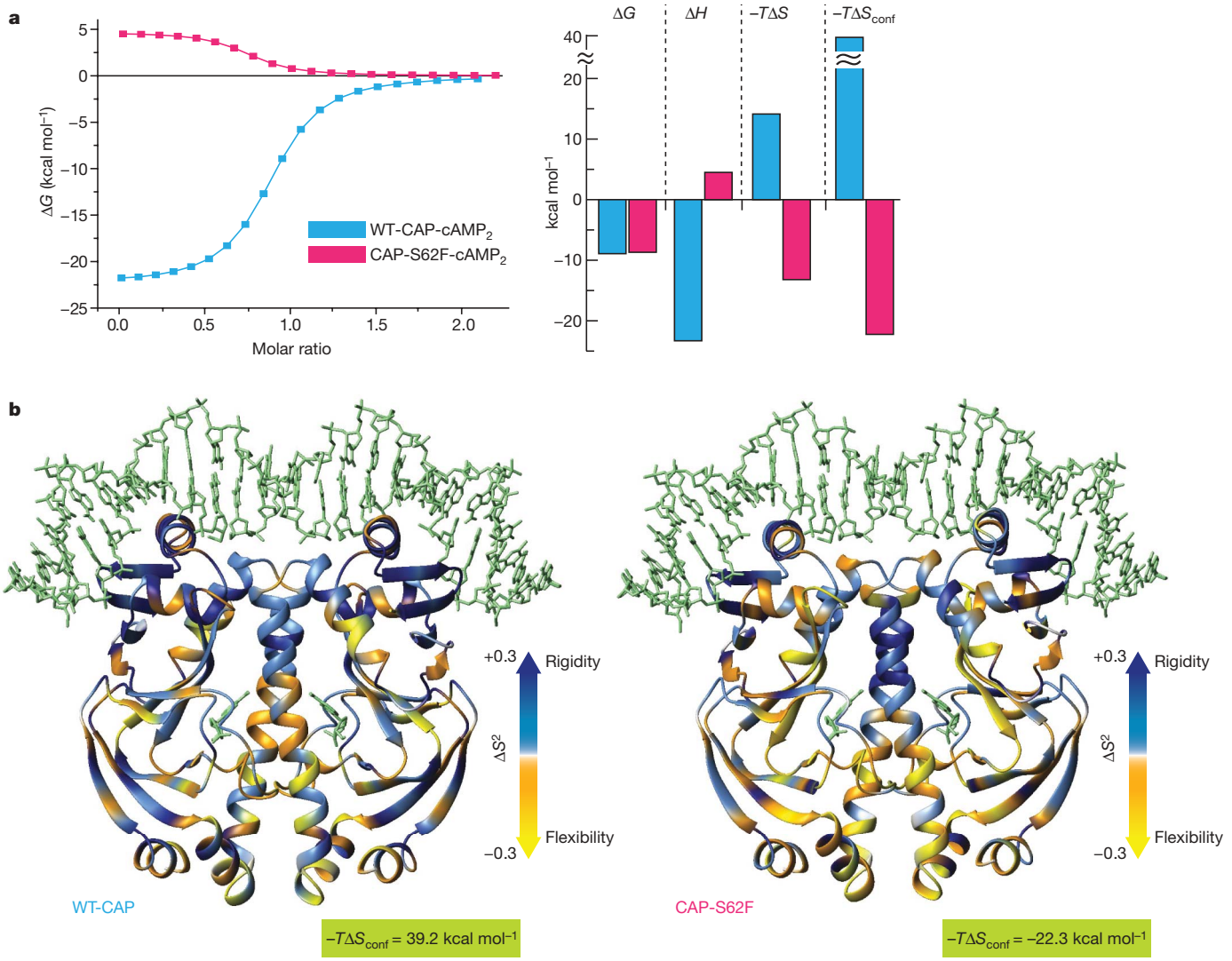


Figure 2 | Energetics of CAP interaction with DNA. a, ITC binding isotherms of the calorimetric titration of a specific DNA sequence to WT-CAP-cAMP$_2$ (blue) and CAP-S62F-cAMP$_2$ (magenta) and the associated thermodynamic components (ΔG, ΔH and ΔS) displayed as bars. $-T\Delta S_{conf}$ is the conformational entropy as measured by NMR. **b**, Effect of DNA binding on N-H bond order parameters of CAP. Changes in order parameters, ΔS^2, for WT-CAP-cAMP$_2$ (left) and CAP-S62F-cAMP$_2$ (right) on DNA binding. ΔS^2 is given as S^2 (after DNA binding) – S^2 (before DNA binding), so positive ΔS^2 values denote enhanced rigidity of the protein backbone on DNA binding. The conformational entropy of DNA complex formation estimated through ΔS^2 values is unfavourable for WT-CAP-cAMP$_2$ ($-T\Delta S_{conf} = 39.2$ kcal mol^{-1}) and favourable for CAP-S62F-cAMP$_2$ ($-T\Delta S_{conf} = -22.3$ kcal mol^{-1}). S^2 plots for all WT-CAP and CAP-S62F liganded states, including error bars, are provided in Supplementary Fig. 13.

S^2 (ΔS^2) are unevenly distributed throughout the protein structure. More specifically, whereas most of the residues in DBD of either WT-CAP or CAP-S62F become more rigid on DNA binding, the majority of the residues that makes up the cAMP-binding site exhibit significantly lower S^2 values in CAP-S62F, indicating an enhancement in motional freedom. It should be noted that, as indicated by chemical shift analysis, DNA binding to CAP-S62F-cAMP$_2$ reorganizes the cAMP-binding pocket (Supplementary Fig. 4b). The relaxation data show that this reorganization results in enhanced flexibility of the cAMP-binding region, presumably by altering the local packing density[21].

Order parameters values are indicative of the amplitude of spatial fluctuations experienced by a bond vector and, thus, can be related to conformational entropy[19]. Despite certain assumptions and limitations, this approach can provide reasonably accurate per-residue entropies[22–27]. By converting order parameters to conformational entropy, $-T\Delta S_{conf}$, we estimate that DNA binding to WT-CAP-cAMP$_2$ is accompanied by an unfavourable conformational entropy change ($-T\Delta S_{conf} \sim 39$ kcal mol^{-1}), whereas DNA binding to CAP-S62F-cAMP$_2$ is accompanied by a favourable conformational entropy change amounting to $-T\Delta S_{conf} \sim -22$ kcal mol^{-1} (Fig. 2b and Supplementary Fig. 13). We conclude that the calorimetrically measured large entropy that drives the strong binding of CAP-S62F-cAMP$_2$ to DNA is dominated by the favourable conformational entropy change (Fig. 2a).

In the case of DNA binding to CAP-S62F-cAMP$_2$ the calorimetrically determined energetics is the sum of two events: the first is the direct binding interaction between DNA and the protein molecules in the active conformation, and the second is the population shift from the inactive to the active conformation (Fig. 4). The thermodynamics of the allosteric transition that results in activation can be extracted by using CAP*-G141S, a constitutively active mutant[9,28] (Supplementary Fig. 14), wherein DBD, as NMR spectra indeed demonstrate (Supplementary Fig. 14a), adopts largely the active conformation in the absence of cAMP. cAMP binds to CAP*-G141S with two orders of magnitude higher affinity than to WT-CAP protein (Supplementary Fig. 14b). This energy difference corresponds to the energy spent by cAMP binding to elicit the active conformation to the wild-type protein. The combined data (Supplementary Fig. 14a–c) indicate that the allosteric transition of DBD from the inactive to the active conformation requires ~ 3.0 kcal mol^{-1} (ΔG_{activ}), with the process being entropy driven ($-T\Delta S_{activ} = -2.2$ kcal mol^{-1}) but enthalpically opposed

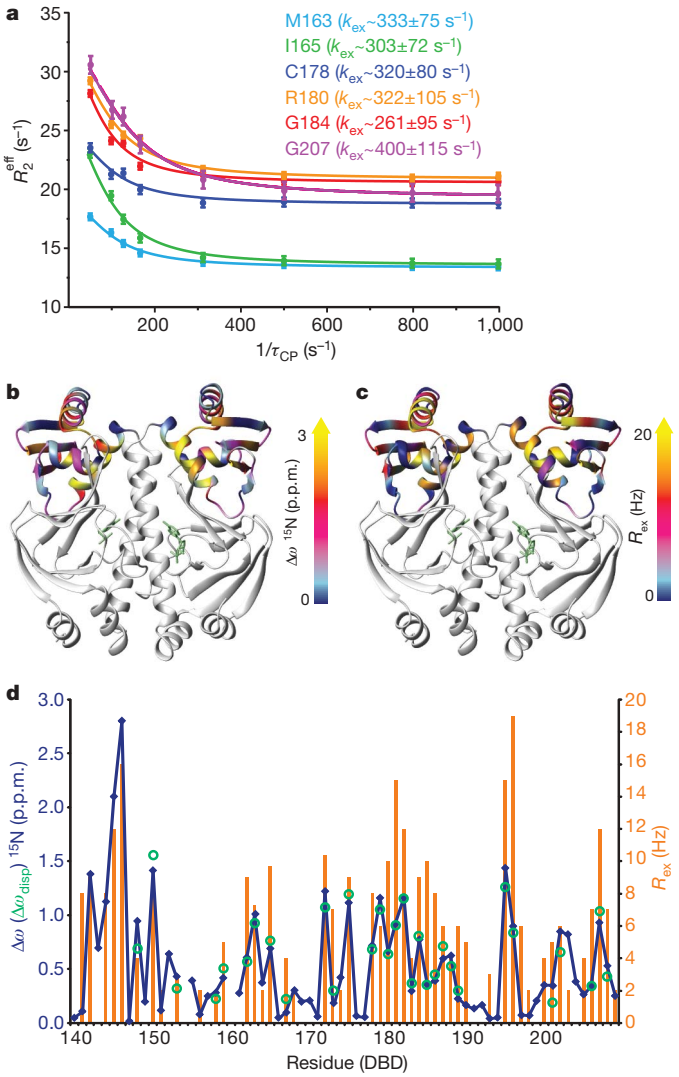


Figure 3 | CAP-S62F-cAMP$_2$ visits an active, low-populated conformational state. **a**, Representative relaxation dispersion data of ^{15}N backbone amides of CAP-S62F-cAMP$_2$ DBD residues. Data were collected and analysed at two external fields (^{1}H 600 and 800 MHz; see Methods), but only the 600 MHz data are shown for clarity. **b**, Backbone ^{15}N chemical shift differences, $\Delta\omega$, of the DBD residues between the apo and cAMP$_2$-bound WT-CAP mapped on the structure. **c**, Conformational exchange dynamics of CAP-S62F-cAMP$_2$ on the μs–ms timescale, as indicated by R_{ex}, are mapped on the structure. **d**, Plot of ^{15}N $\Delta\omega$ values (blue), R_{ex} values (orange), and $\Delta\omega_{disp}$ (green) values between the major and minor conformations in CAP-S62F-cAMP$_2$. The very close correspondence between $\Delta\omega$ (blue rectangles) and $\Delta\omega_{disp}$ (green open circles) values provide strong evidence that the excited conformational state transiently populated by the DBD in CAP-S62F-cAMP$_2$ is the active, DNA-binding competent state.

($\Delta H_{activ} = 5.2$ kcal mol^{-1}). Taken all thermodynamic data together (Fig. 4 and Supplementary Fig. 14), it can be concluded that the large favourable entropy change measured for the formation of the CAP-S62F-cAMP$_2$-DNA complex (–13.2 kcal mol^{-1}; Fig. 2a) is dominated by the direct DNA binding to the protein (–11 kcal mol^{-1}; Fig. 4b).

In conclusion, we present here the intriguing case of an allosteric regulatory protein system that adopts a structurally inactive conformation but is activated for ligand (DNA) binding by favourable changes in conformational entropy originating in protein motions adaptation (Supplementary Fig. 15). The emerging view of protein flexibility as a major source of conformational entropy that can influence the thermodynamics of binding[29,30] indicates that dynamic activation of protein function may be more common than thought.

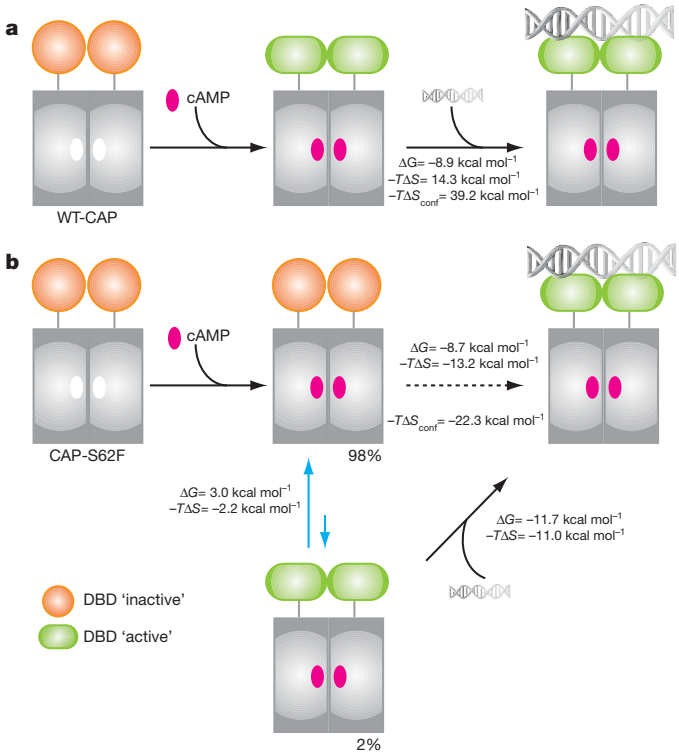


Figure 4 | Reaction pathways and energetics for cAMP-mediated CAP activation and DNA binding. **a**, cAMP binding to WT-CAP elicits the active conformation so that DBD becomes structurally poised to interact favourably with DNA. Complex formation is strongly enthalpically favoured and entropically unfavourable. **b**, cAMP binding to CAP-S62F, in contrast to WT-CAP, stabilizes only marginally the active conformation, which is poorly populated (~2%). Because DNA binds to the active conformation of CAP with many orders of magnitude stronger affinity than to the inactive conformation, DNA will bind selectively to the active, low-populated DBD state and shift the population from the inactive DBD to the active DBD conformation. DNA binding to CAP-S62F-cAMP$_2$ is entirely driven by entropy, which is dominated by conformational entropy. The blue arrows indicate the allosteric transition undergone by DBD. Thermodynamic values were extracted as described in Supplementary Fig. 14.

METHODS SUMMARY

The plasmids expressing for CAP, CAP-S62F and CAP*-G141S, with or without a carboxy-terminally fused histidine tag, were transformed into *Escherichia coli* BL21 (DE3) cells. Isotopically labelled samples were prepared as described in the Methods. NMR experiments were performed on Varian 800- and 600-MHz spectrometers at 32 °C. Backbone assignment and relaxation experiments were performed using standard pulse sequences. Relaxation data were analysed as described in Methods. ITC experiments were performed on an iTC200 microcalorimeter (MicroCal).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 June; accepted 7 October 2009.

1. Kuriyan, J. & Eisenberg, D. The origin of protein interactions and allostery in colocalization. *Nature* **450**, 983–990 (2007).
2. Goodey, N. M. & Benkovic, S. J. Allosteric regulation and catalysis emerge via a common route. *Nature Chem. Biol.* **4**, 474–482 (2008).
3. Smock, R. G. & Gierasch, L. M. Sending signals dynamically. *Science* **324**, 198–203 (2009).
4. del Sol, A., Tsai, C. J., Ma, B. & Nussinov, R. The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure* **17**, 1042–1050 (2009).
5. Lee, J. *et al.* Surface sites for engineering allosteric control in proteins. *Science* **322**, 438–442 (2008).
6. Changeux, J. P. & Edelstein, S. J. Allosteric mechanisms of signal transduction. *Science* **308**, 1424–1428 (2005).
7. Won, H. S., Lee, Y. S., Lee, S. H. & Lee, B. J. Structural overview on the allosteric activation of cyclic AMP receptor protein. *Biochim. Biophys. Acta* **1794**, 1299–1308 (2009).

8. Schultz, S. C., Shields, G. C. & Steitz, T. A. Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science* **253,** 1001–1007 (1991).
9. Popovych, N., Tzeng, S. R., Tonelli, M., Ebright, R. H. & Kalodimos, C. G. Structural basis for cAMP-mediated allosteric control of the catabolite activator protein. *Proc. Natl Acad. Sci. USA* **106,** 6927–6932 (2009).
10. Passner, J. M., Schultz, S. C. & Steitz, T. A. Modeling the cAMP-induced allosteric transition using the crystal structure of CAP-cAMP at 2.1 Å resolution. *J. Mol. Biol.* **304,** 847–859 (2000).
11. Berg, J. M., Tymoczko, J. L. & Stryer, L. *Biochemistry* 6th edn (Freeman, 2006).
12. Dai, J., Lin, S. H., Kemmis, C., Chin, A. J. & Lee, J. C. Interplay between site-specific mutations and cyclic nucleotides in modulating DNA recognition by *Escherichia coli* cyclic AMP receptor protein. *Biochemistry* **43,** 8901–8910 (2004).
13. Baichoo, N. & Heyduk, T. Mapping conformational changes in a protein: application of a protein footprinting technique to cAMP-induced conformational changes in cAMP receptor protein. *Biochemistry* **36,** 10830–10836 (1997).
14. Aiba, H., Nakamura, T., Mitani, H. & Mori, H. Mutations that alter the allosteric nature of cAMP receptor protein of *Escherichia coli*. *EMBO J.* **4,** 3329–3332 (1985).
15. Mittermaier, A. & Kay, L. E. New tools provide new insights in NMR studies of protein dynamics. *Science* **312,** 224–228 (2006).
16. Palmer, A. G. III. NMR characterization of the dynamics of biomacromolecules. *Chem. Rev.* **104,** 3623–3640 (2004).
17. Kern, D. & Zuiderweg, E. R. The role of dynamics in allosteric regulation. *Curr. Opin. Struct. Biol.* **13,** 748–757 (2003).
18. Akke, M., Bruschweiler, R. & Palmer, A. G. III. NMR order parameters and free energy: an analytical approach and its application to cooperative Ca^{2+} binding by calbindin D_{9k}. *J. Am. Chem. Soc.* **115,** 9832–9833 (1993).
19. Yang, D. & Kay, L. E. Contributions to conformational entropy arising from bond vector fluctuations measured from NMR-derived order parameters: application to protein folding. *J. Mol. Biol.* **263,** 369–382 (1996).
20. Cavanagh, J. & Akke, M. May the driving force be with you — whatever it is. *Nature Struct. Biol.* **7,** 11–13 (2000).
21. Zhang, F. & Bruschweiler, R. Contact model for the prediction of NMR N-H order parameters in globular proteins. *J. Am. Chem. Soc.* **124,** 12654–12655 (2002).
22. Kay, L. E., Muhandiram, D. R., Wolf, G., Shoelson, S. E. & Forman-Kay, J. D. Correlation between binding and dynamics at SH2 domain interfaces. *Nature Struct. Biol.* **5,** 156–163 (1998).
23. Bracken, C., Carr, P. A., Cavanagh, J. & Palmer, A. G. III. Temperature dependence of intramolecular dynamics of the basic leucine zipper of GCN4: implications for the entropy of association with DNA. *J. Mol. Biol.* **285,** 2133–2146 (1999).
24. Mauldin, R. V., Carroll, M. J. & Lee, A. L. Dynamic dysfunction in dihydrofolate reductase results from antifolate drug binding: modulation of dynamics within a structural state. *Structure* **17,** 386–394 (2009).
25. Frederick, K. K., Marlow, M. S., Valentine, K. G. & Wand, A. J. Conformational entropy in molecular recognition by proteins. *Nature* **448,** 325–329 (2007).
26. Popovych, N., Sun, S., Ebright, R. H. & Kalodimos, C. G. Dynamically driven protein allostery. *Nature Struct. Mol. Biol.* **13,** 831–838 (2006).
27. MacRaild, C. A., Daranas, A. H., Bronowska, A. & Homans, S. W. Global changes in local protein dynamics reduce the entropic cost of carbohydrate binding in the arabinose-binding protein. *J. Mol. Biol.* **368,** 822–832 (2007).
28. Kim, J., Adhya, S. & Garges, S. Allosteric changes in the cAMP receptor protein of *Escherichia coli*: hinge reorientation. *Proc. Natl Acad. Sci. USA* **89,** 9700–9704 (1992).
29. Wand, A. J. Dynamic activation of protein function: a view emerging from NMR spectroscopy. *Nature Struct. Biol.* **8,** 926–931 (2001).
30. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450,** 964–972 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to L. Kay for critical reading of the manuscript and valuable suggestions. We thank R. H. Ebright and Y. Ebright for providing the DNA fragment and N. Popovych for her help with the preparation of some CAP mutants. This work was supported by National Science Foundation (NSF) grant MCB618259 to C.G.K.

Author Contributions C.G.K. conceived the project. S.-R.T. and C.G.K. designed the experiments. S.-R.T. performed all experiments. S.-R.T. and C.G.K. analysed and interpreted data and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.G.K. (babis@rutgers.edu).

METHODS

Sample preparation. The plasmid expressing for CAP, with or without a C-terminally fused histidine tag, were transformed into *Escherichia coli* BL21 (DE3) cells. CAP mutants (D53H, S62F, G141S, R142H/A144T, L148R) were constructed using site-directed mutagenesis. Cells were grown at 37 °C in the presence of ampicillin. U-^{2}H/^{13}C/^{15}N-labelled CAP was prepared by growing cells in minimal medium in 99.9% $^{2}H_2O$ supplemented with $^{15}NH_4Cl$ (1 g l^{-1}) and $^{2}H_7$/$^{13}C_6$ glucose (2 g l^{-1}) as the sole nitrogen and carbon sources, respectively. Protein synthesis was induced by the addition of 1 mM of isopropyl-β-D-thiogalactoside (IPTG) at D_{600} ~0.4 and cells were harvested after 8–10 h. Cells were re-suspended in lysis buffer containing 20 mM Tris-HCl (pH 8.0), 500 mM KCl and 2 mM β-mercaptoethanol, lysed with a high-pressure homogenizer and centrifuged at 50,000*g*. Two purification steps were used for all protein samples. The first one involved a nickel-chelated Sepharose fast-flow resin (GE Healthcare) and the second one a Superdex-200 size exclusion column (GE Healthcare). Protein concentration was determined spectrophotometrically at 278 nm using an extinction coefficient of 40,800 M^{-1} cm^{-1}. cAMP concentration was determined spectrophotometrically using a molecular extinction coefficient of 14,650 M^{-1} cm^{-1} at 259 nm. A 30-base-pairs-long DNA duplex was used containing the consensus sequence for CAP[31], using a 13-nucleotide (5′-CGAAAAATGTGAT-3′) and a 17-nucleotide (5′-CTAGATCACATTTTT CG-3′) DNA oligonucleotides. The annealed DNA duplex contains symmetry-related single-phosphate gaps between positions 9 and 10 and 13′ and 14′. The DNA duplex was annealed and purified by using ion-exchange chromatography (Mono Q; GE Healthcare).

NMR spectroscopy. NMR experiments were performed on Varian 800- and 600-MHz spectrometers at 32 °C. Protein concentration for samples used for NMR studies was typically 0.4–0.8 mM in 50 mM sodium phosphate (pH 6.0), 500 mM KCl, 1 mM β-mercaptoethanol and 7% (v/v) $^{2}H_2O$. Sequential assignment of the ^{1}H, ^{13}C and ^{15}N protein backbone chemical shifts for CAP was achieved by means of through-bond heteronuclear scalar correlations using three-dimensional (3D) TROSY-based triple resonance sequences. Resonance assignment for CAP-cAMP$_2$ and CAP-cAMP$_2$-DNA complexes was completed by recording a series of two-dimensional (2D) or 3D TROSY HN(CO) spectra on samples wherein the C′ of selected amino acids (Val, Ile, Leu, Ala, Glu and Gly) was specifically ^{13}C-labelled in an otherwise uniformly ^{2}H/^{15}N-labelled background[32]. Data were processed by NMRPipe[33] and analysed with NMRView[34]. The combined chemical shift change of a particular residue on ligand binding was calculated as:

$$\Delta\delta = \sqrt{(\omega_{HN}\Delta\delta_{HN})^2 + (\omega_N\Delta\delta_N)^2 + (\omega_{C\alpha}\Delta\delta_{C\alpha})^2 + (\omega_{C\beta}\Delta\delta_{C\beta})^2 + (\omega_{CO}\Delta\delta_{CO})^2}$$

where ω_i denotes the weight factor of nucleus i; $\omega_{HN} = 1$, $\omega_N = 0.154$, $\omega_{C\alpha} = \omega_{C\beta} = 0.276$ and $\omega_{CO} = 0.341$ (ref. 35).

Relaxation measurement and analysis. Three relaxation parameters were measured for all backbone amides of WT-CAP and CAP-S62F in the cAMP$_2$-bound and cAMP$_2$-DNA-bound states: the ^{1}H-^{15}N nuclear Overhauser effect (NOE) the longitudinal relaxation rate R_1 and the transverse relaxation rate R_2 (ref. 15). ^{15}N R_1 values were measured from 2D spectra recorded with relaxation delays 100, 400, 600, 800, 1,000, 1,300, 1,600, 2,000 and 2,400 ms; ^{15}N R_2 values were measured from 2D spectra recorded with relaxation delays 7.7, 15.5, 23.2, 31.0, 46.6, 54.3, 62.1 and 77.6 ms, respectively. Data sets were acquired as 256 × 1,024 complex points in the $t1 \times t2$ time-domain dimensions. Data points were fitted as a function of the length of the parametric relaxation delay to two-parameter decay curves of the form $I(t) = I_{0e}{}^{-Rt}$, where I is the intensity of the magnetization. ^{1}H-^{15}N NOE data were obtained by recording, in an interleaved manner, one spectrum with a 8 s recycle delay followed by a 8 s saturation and another spectrum with no saturation and a 16 s recycle delay.

ModelFree[36], Fast-ModelFree[37] and Relax[38] were used to optimize the fit of both internal dynamics and global tumbling parameters using the model-free approach[39]. Initial estimates for the rotational diffusion tensor of all complexes were obtained from the ratio of longitudinal and transverse relaxation rates[40]. Residues with ^{1}H-^{15}N NOE value less than 0.65 or with R_1 or R_2 values exceeding one standard deviation from the mean and residues experiencing R_{ex} contributions, as assessed by Carr–Purcell–Meiboom–Gill (CPMG) experiments, were excluded from the fitting[41]. Overall correlation times and rotational diffusion tensors for isotropic, axially symmetric and fully anisotropic models were estimated from the R_2/R_1 ratios of the remaining backbone amide groups and the structures of CAP in the different liganded states using the program quadric_diffusion[42] and TENSOR[43]. The most appropriate diffusion tensor was selected by a comparison of χ^2 goodness-of-fit parameters and using F-statistical analysis. Using these rotational diffusion tensors, backbone relaxation data were fit to the five standard Lipari–Szabo model-free formalism models[44]. The fitted dynamics parameters for each model are as follows: model 1, order parameter (S^2); model 2, S^2 and internal correlation time (τ_e); model 3, S^2 and R_{ex}; model 4, S^2, τ_e and R_{ex}; model 5, order parameters for two time scales ($S_f{}^2$ and $S_s{}^2$) and τ_e for the slower time scale. Parameters of the model-free formalism were optimized for each residue individually, and the best parameter set identified by model selection. All parameters, including the diffusion tensor, were then optimized. This process was repeated until the solution converged. The quality of the fits between the experimental data and each model were calculated as χ^2 statistics, and the different models were then compared to each other using F statistics. Model selection based on these two statistics was performed according to the flowchart outlined previously[44]. Confirmation that the internal dynamic parameters were well optimized was further established by comparing model-free fitted R_{ex} values with those experimentally measured from CPMG experiments. Using the CPMG-measured R_{ex} values to correct the initial R_2 rates eliminated the need for an R_{ex} term to fit the data for the majority of the residues. For both WT-CAP-cAMP$_2$-DNA and CAP-S62F-cAMP$_2$-DNA complexes (τ_C~32.5 ns for both complexes) an isotropic diffusion tensor produced an optimal fit, whereas an axially symmetric diffusion tensor was optimal for the WT-CAP-cAMP$_2$ ($D_{II}/D_{\perp}$ ~1.17; τ_C~22 ns) and CAP-S62F-cAMP$_2$ ($D_{II}/D_{\perp}$ ~1.26; τ_C~22 ns) complexes.

To experimentally characterize and quantify exchange contributions (R_{ex}) for the CAP-S62F-cAMP$_2$ complex, an ^{15}N-TROSY-CPMG pulse sequence was used[45,46] and R_2 relaxation rates were measured at spectrometer frequencies of 600 and 800 MHz. The R_2 relaxation dispersion data were fit simultaneously at the two frequencies using the full Carver–Richards equation[47] for exchange between two sites (A and B), as shown before[46,48–51]

$$R_2^{eff} = \frac{1}{2}\left[R_{2A}^0 + R_{2B}^0 + k_{ex} - \frac{1}{\tau_{CP}}\cosh^{-1}[D_+\cosh(\eta_+) - D_-\cosh(\eta_-)]\right]$$

in which

$$D_{\pm} = \frac{1}{2}\left[\pm 1 + \frac{\psi + 2\Delta\omega^2}{(\psi + \zeta^2)^{1/2}}\right]^{1/2}$$

$$\eta_{\pm} = \frac{\tau_{CP}}{\sqrt{2}}\left[\pm\psi + \left(\psi^2 + \zeta^2\right)^{1/2}\right]^{1/2}$$

$$\psi = \left(R_{2A}^0 - R_{2B}^0 - p_A k_{ex} + p_B k_{ex}\right)^2 - \Delta\omega^2 + 4p_A p_B k_{ex}^2$$

$$\zeta = 2\Delta\omega\left(R_{2A}^0 - R_{2B}^0 - p_A k_{ex} + p_B k_{ex}\right)$$

where τ_{CP} is the time between successive 180° pulses in the CPMG pulse train, $R_{2A(B)}{}^0$ is the exchange-free R_2 relaxation rate of site A(B), p_A and p_B are the populations of states A and B, respectively, k_{ex} is the exchange rate constant and $\Delta\omega$ is the chemical shift difference between states A and B. Data were analysed using the program CPMGFit[52] and CPMG_fit provided by L. Kay. The range of k_{ex} values for the vast majority of DBD residues is 260–400 s^{-1} (Fig. 3a). The relatively slow exchange process, together with the relatively large $\Delta\omega$ values between the two states for many DBD residues ($\Delta\omega$~300–1,000 rad s^{-1}), render the exchange process for these residues slow-to-intermediate on the NMR chemical shift time scale. This conclusion is consistent with the proportionality factor α values[53] for these residues ($\alpha \leq 1$), measured at spectrometer frequencies of 600 and 800 MHz. The similarity of exchange rates extracted from fits of dispersion profiles on a per-residue basis for the DBD region supports indeed a two-site model of exchange between the inactive and active DBD conformation. The excellent correlation between $\Delta\omega$ calculated from CPMG experiments and $\Delta\omega$ measured for the apo-WT-CAP and WT-CAP-cAMP$_2$ states (Fig. 3d) further corroborates this conclusion. To exclude contributions from protein aggregation R_2 relaxation rates were also measured at lower sample concentrations (~0.1 mM).

Assessment of conformational entropy from order parameters. The contribution of protein motions to the conformational entropy ($-T\Delta S_{conf}$) of DNA binding to WT-CAP-cAMP$_2$ and CAP-S62F-cAMP$_2$ was estimated by measuring changes in the order parameter, S^2, as a function of the ligation state, as described above. The conformational entropy of the CAP backbone is then calculated by[19]:

$$\Delta S = -k_B\sum_i \ln\left\{\frac{3-(1+8S_{a,i})^{1/2}}{3-(1+8S_{b,i})^{1/2}}\right\}$$

in which k_B is the Boltzmann constant, and S_a and S_b are the order parameters for state a (cAMP$_2$-bound) and state b (cAMP$_2$-DNA-bound).

ITC experiments. Calorimetric titrations of CAP with cAMP and/or DNA were performed on an iTC200 microcalorimeter (MicroCal) at 32 °C. Protein samples were extensively dialysed against the ITC buffer containing 50 mM KPi (pH 6.0), 500 mM KCl and 1 mM Tris(2-carboxyethyl)phosphine (TCEP). All solutions were filtered using membrane filters (pore size 0.45 μm and thoroughly degassed

for 20 min by gentle stirring under vacuum. The 200-μl sample cell was filled typically with a 5–50 μM solution of protein and the 40-μl injection syringe with 600 μM of the titrating cAMP or 50–500 μM of the titrating DNA. The ligand solution was prepared by dissolving cAMP or DNA in the flow through of the last buffer exchange. Each titration typically consisted of a preliminary 0.2-μl injection followed by 18 subsequent 2-μl injections. Data for the preliminary injection, which are affected by diffusion of the solution from and into the injection syringe during the initial equilibration period, were discarded. Binding isotherms were generated by plotting heats of reaction normalized by the modes of injectant versus the ratio of total injectant to total protein per injection. The data were fitted using the sequential binding site model embedded in Origin 7.0 (MicroCal).

31. Parkinson, G. *et al.* Structure of the CAP-DNA complex at 2.5 Å resolution: a complete picture of the protein-DNA interface. *J. Mol. Biol.* **260,** 395–408 (1996).
32. Takeuchi, K., Ng, E., Malia, T. J. & Wagner, G. 1-^{13}C amino acid selective labeling in a ^{2}H^{15}N background for NMR studies of large proteins. *J. Biomol. NMR* **38,** 89–98 (2007).
33. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–293 (1995).
34. Johnson, B. A. Using NMRView to visualize and analyze the NMR spectra of macromolecules. *Methods Mol. Biol.* **278,** 313–352 (2004).
35. Evenäs, J. *et al.* Ligand-induced structural changes to maltodextrin-binding protein as studied by solution NMR spectroscopy. *J. Mol. Biol.* **309,** 961–974 (2001).
36. Palmer, A. G. III. ModelFree. ⟨http://www.palmer.hs.columbia.edu/software/modelfree.html⟩.
37. Cole, R. & Loria, J. P. FAST-Modelfree: a program for rapid automated analysis of solution NMR spin-relaxation data. *J. Biomol. NMR* **26,** 203–213 (2003).
38. d'Auvergne, E. J. & Gooley, P. R. Optimisation of NMR dynamic models I. Minimisation algorithms and their performance within the model-free and Brownian rotational diffusion spaces. *J. Biomol. NMR* **40,** 107–119 (2008).
39. Lipari, G. & Szabo, A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc.* **104,** 4546–4559 (1982).
40. Tjandra, N., Feller, S. E., Pastor, R. W. & Bax, A. Rotational diffusion anisotropy of human ubiquitin from ^{15}N NMR relaxation. *J. Am. Chem. Soc.* **117,** 12562–12566 (1995).
41. Hwang, P. M., Skrynnikov, N. R. & Kay, L. E. Domain orientation in beta-cyclodextrin-loaded maltose binding protein: diffusion anisotropy measurements confirm the results of a dipolar coupling study. *J. Biomol. NMR* **20,** 83–88 (2001).
42. Palmer, A. G. III. quadric_diffusion. ⟨http://www.palmer.hs.columbia.edu/software/quadric.html⟩.
43. Dosset, P., Hus, J. C., Blackledge, M. & Marion, D. Efficient analysis of macromolecular rotational diffusion from heteronuclear relaxation data. *J. Biomol. NMR* **16,** 23–28 (2000).
44. Mandel, A. M., Akke, M. & Palmer, A. G. III. Backbone dynamics of *Escherichia coli* ribonuclease HI: correlations with structure and function in an active enzyme. *J. Mol. Biol.* **246,** 144–163 (1995).
45. Loria, J. P., Rance, M. & Palmer, A. G. III. A TROSY CPMG sequence for characterizing chemical exchange in large proteins. *J. Biomol. NMR* **15,** 151–155 (1999).
46. Mulder, F. A., Mittermaier, A., Hon, B., Dahlquist, F. W. & Kay, L. E. Studying excited states of proteins by NMR spectroscopy. *Nature Struct. Biol.* **8,** 932–935 (2001).
47. Carver, J. P. & Richards, R. E. A general two-site solution for the chemical exchange produced dependence of T2 upon the Carr-Purcell pulse separation. *J. Magn. Reson.* **6,** 89–105 (1972).
48. Watt, E. D., Shimada, H., Kovrigin, E. L. & Loria, J. P. The mechanism of rate-limiting motions in enzyme function. *Proc. Natl Acad. Sci. USA* **104,** 11981–11986 (2007).
49. Henzler-Wildman, K. A. *et al.* Intrinsic motions along an enzymatic reaction trajectory. *Nature* **450,** 838–844 (2007).
50. Boehr, D. D., McElheny, D., Dyson, H. J. & Wright, P. E. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* **313,** 1638–1642 (2006).
51. Korzhnev, D. M. *et al.* Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR. *Nature* **430,** 586–590 (2004).
52. Palmer, A. G. III. CPMGFit. ⟨http://www.cumc.columbia.edu/dept/gsas/biochem/labs/palmer/software/cpmgfit.html⟩.
53. Millet, O., Loria, J. P., Kroenke, C. D., Pons, M. & Palmer, A. G. The static magnetic field dependence of chemical exchange linebroadening defines the NMR chemical shift time scale. *J. Am. Chem. Soc.* **122,** 2867–2877 (2000).

NEWS

Emerging shortages

The booming economies of Brazil, China, India, Singapore and possibly Russia could face significant science and engineering workforce shortages as soon as next year, a group of recently released studies suggests.

Demographers from these countries analysed government data to predict the future size of the science and engineering workforce, as part of a project called 'The Evolving Global Talent Pool', coordinated by the State University of New York's Levin Institute in New York City. They presented their analyses at a 30 October symposium in Manhattan (see go.nature.com/9gAO2c). Although the various analyses were released at different times over the past four years, this is the first time that country trend comparisons have been highlighted.

The numbers suggest that mismatches between talent supply (the number of graduates that the countries' universities are cranking out) and demand in government, industry and academic sectors could become particularly acute within a decade. According to the report, significant shortages of graduates are likely to be evident by next year, despite considerable investment in higher education and research and development by each country's government. This raises a serious question, it says, about whether corporations, education systems and societies understand the demand issue correctly.

China's supply of relevant professionals, for example, is expected to be 3.48 million next year. But demand will have soared to some 3.85 million by then, and will rise to 5.9 million in 2015, according to data compiled by Denis Fred Simon, a professor at Pennsylvania State University's School of International Affairs in University Park, and Cong Cao, a senior research associate at the Levin Institute.

India faces a shortage of 60,000 engineers by 2010 and 2.45 million by 2020, according to R. Venkatesan and Wilima Wadhwa of India's National Council of Applied Economic Research in New Delhi. And Poh Kam Wong of the National University of Singapore predicted that the island state would produce just 55.2% of the science and technology professionals it would need in 2010, requiring some 44,000 foreigners to fill the gap.

However, although there are big shortages in some fields in these countries, there are surpluses in others. Aspiring scientists and researchers may not receive market signals about job availability quickly enough, according to Simon.

"Very few countries do demand-side analysis," says Simon. A lack of quality training and education also contributes to the shortages, says the report, because many graduates in these countries have degrees from institutions with a high proportion of poorly qualified faculty.

The studies have an important caveat: the type and extent of the data available in each country vary widely, and governments use different demographic categorizations of science and technology professionals. China's statistics, for example, included as science or engineering professionals people without degrees who have been working in those fields for at least 10 years. ■

Gene Russo

POSTDOC JOURNAL

Climate-change depression

My freelancing has kept me busy. I have multiple projects due by the end of the month, focusing mostly on climate change and biodiversity; the next few weeks are going to be hectic. Seasoned freelancers tell me this is the way of my new world. It's feast or famine.

Trying to work from home during my son's naps can be a struggle. More challenging yet is trying not to get sucked into a vortex of depression regarding the potentially apocalyptic implications of the topics I'm writing about.

Apparently I'm not the only one fighting this condition. It even has its own name, 'climate-related depression'. And Australian doctors recently reported a case of 'climate-change delusion' in which a teenager refused to drink water, as he was convinced that he would cause a major water shortage in drought-stricken Australia. Extinctions here, range shifts there, and pathogens and pests expanding everywhere. With all the tabloid reports, it is easy to become haunted by nightmares of a stark Earth devoid of life.

But it's not all bad news, and to keep my sanity I focus on the positives. Animals and plants have an amazing capacity to adapt. We can help them. Many of my projects aim to help educate and inform. And let's face it: my freelance work, much of it catering to academics and policymakers, will have a far greater impact than most of the scientific-journal papers from my postdoc days. I believe that what I write matters. That is my cure for climate-related depression. ■

Joanne Isaac was a postdoc in climate-change effects on biodiversity at James Cook University, Townsville, Australia.

IN BRIEF

Postdocs on a pittance

Postdocs in Canada are underpaid and face uncertain career prospects, according to a 3 November report by the Canadian Association of Postdoctoral Scholars. *A Postdoctoral Crisis in Canada* says that current postdoc stipends average Can$30,000 (US$28,600), less than a graduate student scholarship, which it says averages $45,000, and less than the $37,000 pay of an entry-level research technician. Based on a survey conducted earlier this year, the report suggests that Canadian postdocs are less likely to take a university post than two decades ago, in part because there are fewer such jobs. Some 55% said they were pleased with their role and training as a postdoc.

Tenure or family?

Marriage and childbirth are what stop most female US graduate students from becoming tenured researchers, according to a report by Washington DC think tank the Center for American Progress (CAP) and the University of California, Berkeley, School of Law. *Staying Competitive: Patching America's Leaky Pipeline in the Sciences* found that married mothers with a PhD are 35% less likely to enter a tenure-track position in the sciences than married fathers with PhDs, according to a National Science Foundation survey. And they are 27% less likely than their male counterparts to get tenure after securing a tenure-track post. The report advises universities and funding agencies to create family-friendly policies, including six weeks of paid maternity leave and a week of paid parental leave.

Huge cuts by drug firms

Pfizer is closing 35% of its global research and development space, according to a 9 November announcement. The New York-based drug company, which employs 14,500 people in research and development worldwide, has said that R&D personnel cuts associated with the closures will make up a significant percentage of the 15% company-wide job cuts planned. Pfizer, which last month acquired US drugmaker Wyeth, has disclosed no further information and did not return phone calls by press time. In early November, US drugmaker Johnson & Johnson announced plans to lay off about 8,000, but did not reveal how the cuts would affect its R&D personnel.

Quantum potential

The emerging field of quantum information science is harnessing nature's strangest habits — and providing an academic haven for young physicists. **Eric Hand** reports.

If history is any guide, codes are made to be broken. But not the codes coming through a window and into a detector in Ray LaFlamme's office. LaFlamme, director of the Institute for Quantum Computing (IQC) in Waterloo, Canada, says that these codes, contained in pulses of light, are "unbreakable". But the information is not hidden within a complicated new cipher. Rather, it is protected by the laws of quantum mechanics, which allow the pulses of light to be enshrouded so delicately that any eavesdropping immediately triggers alarms. The technique, called quantum cryptography, will one day eradicate eavesdropping altogether, says LaFlamme.

With efforts like this, LaFlamme's institute is tackling the broad questions of quantum information science (QIS), an exploding field that blends physics, maths and computer science in ways that, just 25 years ago, were unknown or deemed irrelevant. Institutes and governments worldwide are competing for top physics, maths and computer-science talent. More than 100 academic and industry groups worldwide are now publishing in the QIS field, according to a list that LaFlamme gathered in 2008.

"It's really amazing," says Anton Zeilinger, scientific director of the Vienna laboratory of Austria's Institute for Quantum Optics and Quantum Information (IQOQI). "For a younger person, it's a golden age."

Expansion ahead

The field of QIS focuses on efforts to turn the weirdness of quantum mechanics to its advantage in two main areas: information processing and encryption. So some researchers are doggedly trying to build quantum computers, which attempt to use delicate and blurry bits made possible by quantum mechanics to drastically outperform classical computers. And others work on quantum cryptography, which harnesses the idiosyncratic properties of quantum mechanics to create spy-proof systems, such as the one in LaFlamme's office. Jokingly, LaFlamme says that he could be using the system to send his colleagues secret strategy documents for the institute — and no one else would ever know.

Artur Ekert: renowned.

But one element in that strategy is clear: expansion. The IQC will soon move into its fourth building since it was founded in 2002 at the University of Waterloo, Ontario, having reached capacity time and time again. The new Can$160-million (US$151-million) Quantum-Nano Centre, due to be completed by the end of 2010, should finally offer LaFlamme a space with room to grow. It helps to have a benefactor such as Mike Lazaridis, the founder of Research in Motion, the Waterloo-based company that makes BlackBerry smart phones. After founding the Perimeter Institute for Theoretical Physics a decade ago with Can$150 million of his own money, Lazaridis went on to spin off the IQC with nearly Can$50 million more (see go.nature.com/peFJuq). LaFlamme came to Waterloo with a handful of other researchers to establish the IQC, which now has some 110 researchers; the new building will be able to host more than 200.

Across the world

Growth in the QIS field stretches far beyond Canada. Hot spots in Europe include a robust research presence in the United Kingdom at the University of Oxford, where pioneers such as David Deutsch got the field going in the late 1980s. Another important locale is the IQOQI, a venture of the Austrian Academy of Sciences, which has a lab at the University of Innsbruck as well as in Vienna. Ten years ago, fewer than 40 people were involved in the field in Austria, says Zeilinger; there are now more than 200 researchers associated with the IQOQI.

Singapore has signalled an interest in the field with big recruits such as Artur Ekert, a renowned physicist at the University of Oxford. While keeping his Oxford post, Ekert will spend part of his time at the National University of Singapore as director of the Centre for Quantum Technologies, which in 2007 was given start-up funds of S$150 million (US$108 million) for five years. The centre now has some 100 researchers, he says.

Postdocs such as Bill Coish at the IQC see plenty of opportunities. Coish is looking ahead to an academic future that he thinks is bright in spite of the economic downturn.

"I think it's much better for me than for people working in older fields," says Coish. During his PhD in theoretical physics at the University of Basel in Switzerland, he concentrated on understanding 'decoherence', the tendency of prototype quantum computers to dissolve back into classical states. Switzerland is home to one of the few QIS start-up companies: Geneva-based id Quantique, which in 2007 used quantum cryptography to transmit election results. Although Coish is open to returning to Europe, he's eschewing such start-up firms in favour of academic jobs. He thinks that start-ups want people with a penchant for tinkering in the lab. "Experimentalists have much better luck going to a start-up," he says.

As he looks ahead, physicist Coish says he is trying to hone his skills in maths and computer science to make himself more appealing to academic recruiters. "They want someone who can work a little bit in the traditional fields of theoretical physics, and at the same time branch out into areas that have been traditionally studied by computer scientists and mathematicians," he says.

LaFlamme agrees, saying that there is no

GOSALDAR

SPL

set or required background for the three faculty positions he is hoping to fill this year at the IQC. "It is a field that is inherently interdisciplinary," he says. "My advice would be to be exceedingly good, but have a broad mind." At the IQC, he says, about half the researchers have backgrounds in maths and computer science, and the other half come from physics and physical chemistry. However, a 2007 review of European Commission-funded quantum-information projects found that researchers in experimental or theoretical physics dominated the field.

J. GODANY

Anton Zeilinger: golden age.

Breaking the code

The peculiar properties of quantum computing promise groundbreaking computational speed in special situations, such as code-breaking. In 1994, mathematician Peter Shor, now at the Massachusetts Institute of Technology in Cambridge, showed that a quantum computer would be exponentially faster at factoring numbers than a classical computer. This could fundamentally jeopardize the security of information of all sorts of things — from financial transactions to state secrets — because today's encryption protocols typically derive from the difficulty that classical computers have in factoring. Quantum computers, still in their infancy, take advantage of one aspect of quantum mechanics, where particles seem to be in two places at once. A quantum bit, or qubit, can very quickly outperform regular bits; it would take a million classical bits to simulate the memory held in just 20 qubits.

Not surprisingly, this computational potential has piqued the interest of many government agencies. For example, work in the United States has been funded by the National Science Foundation, the National Security Agency and the Intelligence Advanced Research Projects Activity, which was formed in 2007. But Richard Hughes, a physicist at Los Alamos National Laboratory in New Mexico, says that without any agency "owning" this research, US funding agencies have lacked a coherent plan, and that as a result, both research and job prospects are much stronger in Europe.

Although it is still early days for quantum computing, the other major field within QIS, quantum cryptography, is further along. Just as quantum computers create the potential problem of hacked bank transactions, quantum mechanics leads to a solution. In quantum cryptography, information is sent among multiple parties with decoder keys that are quantum-mechanically entangled. These entangled keys — typically in the form of polarized photons — are made so that when they are sent off to detectors, they register in perfect agreement with each other. But if an eavesdropper were to measure either of the photons en route, the entanglement evaporates for both of the photons, as if one photon 'feels' the interference on the other.

Outwitting the eavesdropper

A few start-up firms, such as id Quantique, have made tentative steps at commercializing this encryption technology. However, challenges remain — for instance, in boosting data-transmission rates.

That's the aim of Chris Erven, a physics PhD student at the IQC, as he heads to a building on the Waterloo campus to install improved laser optics in a system that emits entangled photons. Piped to the building's rooftop, they are beamed in two directions. One set heads to the current off-campus home of the IQC, half a kilometre away. Another set streams in the direction of the Perimeter Institute, 1.3 kilometres away, past a tree that needs to be pruned every so often — what Erven calls the "arboreal eavesdropper" in his system.

Erven climbs up to the roof and points to the rising concrete structure of the Quantum-Nano Centre, which will be specially damped to help limit the effect of vibrations on the fragile quantum devices. "It'll be done just in time for me not to be able to use it," says Erven, who is due to finish his PhD next year. "I'm sure everyone else is excited."

But with his field booming, he's not worried about finding another lab to call home. Interested in engineering, maths and physics, Erven earned a bachelor's degree from the University of Waterloo in applied science — but quickly recognized QIS as the way to unite all his interests. He recalls attending a talk by Shor, and was impressed that he could interact with the person whose work helped to kick-start an entire field. "It's a little different from other fields, where the founders have been dead for 300 years," he says. ■

Eric Hand is a reporter for *Nature* based in Washington DC.

A letter from the past

On the dubious position of *Aelfus* in the evolutionary tree of mankind.

Ruy José Válka Alves

Dear Wallace,

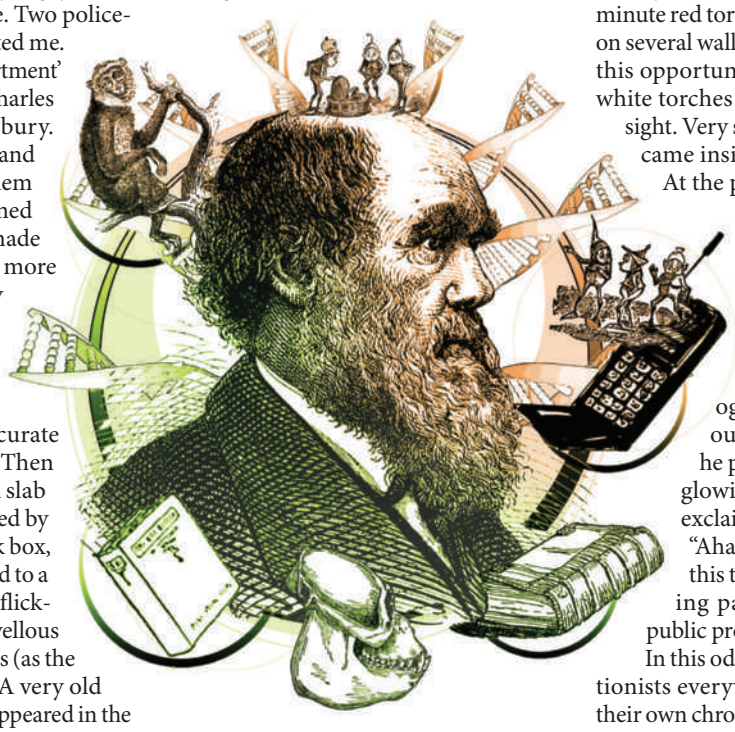
On May 11th 1875, while I was at Down House studying whether the strange mathematics Mendel has been sending me affected corollas of *Impatiens balsaminea*, I fell asleep … I woke in a snow-clad street and was almost run over by several noisy horseless carriages, driven by angry, vehemently gesticulating people. Two policemen with odd accents arrested me.

At the 'Seattle Police Department' I told them my name was Charles Robert Darwin, of Shrewsbury. They all burst into laughter, and became hostile. I showed them my letter of safe conduct signed by Queen Victoria, which made matters even worse for me: more laughter was followed by them insisting on valid travel documents, visa etc. When I demanded to know what a 'photo' was, he showed me a remarkably small and accurate colour painting of himself. Then he tapped onto some black slab upon his table, which was tied by a string to a quite large black box, from which another string led to a frame with a portrait which flickered and had the most marvellous power of changing its colours (as the Octopus from Sao Tiago). A very old man with a full white beard appeared in the frame. "That was Charles Darwin. You are over a century late!" the chief exclaimed. The image was somewhat familiar, but I am much younger!

As I insisted on my name, the chief accused me of being a lunatic, an illegal alien and a liar. He took a small box out of his pocket, and a flickering elf appeared in it. He spoke intermittently, but I could not hear the elf reply. All around the station, several other policemen were also talking to their own little boxes. And who was the true lunatic?

The only food they brought me that I could identify was a banana, and how a ripe one made its way to this temperate location in midwinter remains quite a mystery. They forced me through a back door into one of the horseless carriages. On the streets, I saw many more people talking to their elves. Soon after we arrived at a 'Mental Institution'.

During several weeks I stayed at that horrible place, but one night someone left a window open so I managed to flee. It was cold and the only open door led me into a public library. Having nowhere to go, I wandered inside and found a considerable collection of books on all subjects, and most importantly, a bit of peace to recover from my previous adventures. Not only did no one bother me at all, but a smiling middle-aged lady with thick spectacles even offered me a glass of water (which tasted like iodine). Many works on the shelves were new to me even though their dates were quite old. By far the most surprising were several papers which I am still working on. However, this is the only fact that suggests I travelled forward in time.



JACEY

I examined the library. Trees of life are used everywhere, for multiple purposes. *Darwinism, Design and Public Education* was quite entertaining, even though I failed to understand what RNA and DNA are. And then, quite frankly, 'Darwinism' is an abominable name! I also chanced upon several articles in the periodical *Scientometrics*, which made me shudder: the scientists of this mysterious land and time have nothing better to do than judge the works of their fellows, based on this odd belief that *where* and how frequently you publish is more important than *what* you publish!

Although many nations in Europe report elves, I never believed in them. Maybe I was mistaken about the workings of the origins of Mankind! Searching for a specimen elf in the library, I first broke open one of those small boxes someone had forgotten on the table. Nothing. Then I opened several of the big boxes which had the flickering frames, and searched around with my hand. Something bit me in the hand as a large blue spark charged out of the box. All of a sudden the entire library became pitch dark, and the next minute red torches lit up all by themselves on several walls. The elves must have used this opportunity to escape, as when the white torches lit up again, none were in sight. Very shortly after that, the police came inside and arrested me again.

At the police station, they stuffed parts of the broken boxes into small envelopes made of an incredible material: it was as transparent as clear glass, but more flexible than silk! At first the police chief had not recognized me from the previous incident, but soon after he pressed my thumb against a glowing red box on his table, he exclaimed, with a sarcastic grin: "Aha, Mr 'Daaarwin' again! And this time, you are here for escaping parole, theft and breaking public property!"

In this odd time, there are more creationists everywhere and they even have their own chronicles and institutions; people keep elves in boxes and speak to them in a primitive accent; men of science are judged by where and not what they publish; oh yes, of course: and the potential of horses for transportation has not yet been discovered! Considering these facts, it seems beyond reasonable doubt that I must have travelled backward in time, even though everyone here claims this is the early twenty-first century!

I have failed so far, but I think elves must stand beside modern man at the top of the tree. However, unless I catch a pair once I get out of prison, you shall hear of me no more …

I remain, yrs etc.,

Chas. Darwin. ■

The author is associate professor at the Botanical Department, National Museum, Federal University of Rio de Janeiro, specializing on oceanic islands and mountain floras.

Join the discussion of Futures in *Nature* at go.nature.com/QMAm2a

FUTURES